METHODS TO ADDRESS CONFOUNDING AND HETEROGENEITY IN COST-EFFECTIVENESS ANALYSES USING REAL-WORLD DATA

Author:
Silvia Moler Zapata

Thesis submitted in accordance with the requirements for the degree of Doctor of Philosophy of the University of London

August 2023

Department of Health Services Research and Policy
Faculty of Public Health and Policy
LONDON SCHOOL OF HYGIENE & TROPICAL MEDICINE
UNIVERSITY OF LONDON

Funded by NIHR ARC North Thames

Declaration of Authorship

I, Silvia Moler Zapata, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Signed: Silvia Moler Zapata
Date: 22.08.2023

Abstract

This thesis is concerned with improving methods for cost-effectiveness analyses (CEA). Real-World Data (RWD), for example, from routine data sources such as electronic health records, is used to generate comparative effectiveness and cost-effectiveness evidence in settings where appropriate evidence from Randomised Controlled Trials (RCTs) is not available. However, studies using RWD face fundamental issues pertaining to the study design, in particular around the risk of bias due to confounding and treatment effect heterogeneity. The aim of this thesis is to contribute to the literature on CEA methods for those settings. The thesis considers recent advancements in the causal inference and econometrics literature to examine the following objectives: (i) to identify challenges for comparative- and cost-effectiveness studies in applying the 'target trial' framework, (ii) to evaluate a novel local instrumental variable (LIV) approach in a CEA, (iii) to evaluate the performance of the LIV approach according to varying levels of instrument strength in a simulation study.

The first paper in the thesis considers the main challenges in applying the target trial framework in comparative effectiveness and cost-effectiveness studies that use RWD, and offers recommendations, in particular around the interrelated issues of defining the study population and the comparator groups. The second paper is concerned with methods to address unmeasured confounding and heterogeneity, which are major challenges in CEA that use RWD. In this paper, I evaluate LIV methods in the context of a CEA that uses routine data from the 'Emergency Surgery OR noT' (ESORT) study. In the third paper, I extend this assessment of LIV methods with a simulation study that assesses the performance of LIV in realistic scenarios, defined by varying levels of instrument strength, and different forms of heterogeneity and sample sizes. The findings from these papers suggest that, in addressing both confounding and heterogeneity, LIV methods can provide accurate estimates of treatment effects of direct decision-making relevance. I find that, provided the instrument is strong, or the sample size is at least moderate, the LIV approach reports estimates with low bias and that are statistically efficient, regardless of the form of treatment effect heterogeneity that is present.

The thesis concludes that by directly addressing confounding and heterogeneity the proposed methods can mitigate concerns about studies using RWD. Findings from this thesis can help future CEA that use RWD, to provide more useful evidence for decision-making.

Acknowledgements

First and foremost, I would like to thank my supervisor, Richard Grieve, for the trust and encouragement to pursue this PhD, and for the insightful advice and constant support, which were instrumental for completing it. I am also grateful to Richard Silverwood for his help and feedback in preparation of this thesis.

I have been extremely lucky to work with Stephen O'Neill, David Lugo-Palacios and Andrew Hutchings, who always found the time to meet with me, review my do-files and offer feedback on my work. My gratitude goes to them too.

I would also like to thank Anirban Basu and Luke Keele, as well as the members of the ESORT team, especially Rob Hinchliffe, Neil Smart and David Cromwell, for their valuable contributions to different elements of this thesis.

This thesis was funded by the NIHR ARC North Thames, to whom I am extremely grateful.

I also want to thank my fellow PhD students, who provided much laughter and biscuits, as well as support over video calls, which were eventually replaced by long coffee breaks and chats in the sun.

Finally, I wish to thank my family for their love, patience and unconditional support, they have given me the strength to bring this task to an end; and to my friends for always being there, and for filling with joy the few moments of free time I had.

Contents

Appendices

List of Tables

List of Figures

List of Abbreviations

2SLS – Two-stage least squares

2SRI – Two-stage residual inclusion

ATE – Average treatment effect

CATE – Conditional average treatment effect

CEA – Cost-effectiveness analysis

CI – Confidence interval

CUA – Cost-utility analysis

DALY – Disability-adjusted life year

DAOH – Days alive and out of hospital

EHR – Electronic health records

ES – Emergency surgery

ESORT – Emergency Surgery or not

GDP – Gross Domestic Product

GLM - Generalised Linear Model

HEAP – Health economics analysis plan

HES – Hospital episode statistics

HRQoL – Health-related quality of life

HTA – Health technology assessment

ICD – International classification of diseases

ICER – Incremental cost-effectiveness ratio

ICU – Intensive care unit

INB – Incremental net monetary benefit

ITT – Intention to treat

IV – Instrumental variable

LATE – Local average treatment effect

LIV – Local instrumental variable

MTE – Marginal treatment effect

NELA – National emergency laparotomy audit

NES – Non-emergency surgery

NHS – National health system

NICE – National institute for health and care excellence

ONS – Office for National Statistics

OPCS - Office of population censuses and surveys

PeT – Person centered effects

PP – Per protocol

PS – Propensity score

QALY – Quality-Adjusted Life Year

RCT – Randomised controlled trials

RMSE – Root mean squared error

RWD – Real world data

SA – Sensitivity analysis

SCARF – Secondary care administrative records frailty

SD – Standard deviation

TTE – Target trial emulation

TTO – Tendency-to-operate

Chapter 1. Introduction

1.1 Cost-effectiveness analysis for decision-making in health care

The primary goal of most health systems worldwide is to improve population health (World Health Organization, 2000). In budget-constrained settings, decision-makers and reimbursement agencies have to make difficult decisions about how to assign limited health resources to alternative uses. In many countries, reimbursement agencies draw on evidence from health economic evaluations which compare relative outcomes from alternative interventions, programs or technologies (herein referred to as 'interventions') against their relative costs (i.e., the value of the health forgone elsewhere in the system as a result of the adoption of the intervention) (Drummond et al., 2005).

The most widely adopted form of economic evaluation is cost-effectiveness analysis (CEA), which contrasts the additional cost per additional unit of health outcome, generally defined according to measures such as quality-adjusted life years (QALYs) or disability-adjusted life years (DALYs) (Culyer, 2010).[1] A common metric in CEA is the incremental cost-effectiveness ratio (ICER). The ICER lends itself to a decision rule to judge 'value for money', whereby if the ICER is below a particular cost-effectiveness threshold representing the opportunity cost,[2] then the intervention may be deemed relatively cost-effective and, pending consideration of other issues, such as the level of uncertainty, or the value of innovation, the intervention may be recommended for adoption.[3]

Evidence from CEA can inform resource-allocation decisions in public health and social care, but is most commonly used in recommendations about adoption and use of health interventions in health technology assessment (HTA) (Claxton et al., 2010; Ochalek et al., 2020; Rudrapatna and Butte, 2020; Sorenson et al., 2008). While other factors such as equity may well be important considerations in the development and

[1] This thesis considers CEA in its widest context, so the definition of 'health outcome' adopted includes clinical outcomes but also health-related utility measures.

[2] See Eckermann and Pekarsky (2014) for a review of methods for evaluating a threshold value for the effects of new health interventions.

[3] Alternatively, cost-effectiveness can be assessed looking at the incremental net monetary benefit (INB), which is the difference between the incremental costs and the incremental health outcomes, expressed in monetary terms.

formulation of HTA guidelines, cost-effectiveness has been shown to be the main determinant of previous HTA decisions by the National Institute for Health and Care Excellence (NICE) in England and Wales (Dakin et al., 2015).

Given the central role of economic evidence in HTA decision-making, it is important that CEA use robust analytical methods for assessing the relative effectiveness and costs of health interventions. To that end, NICE publishes its own methods and processes manuals which set out the type of evidence required and how to ensure it is of ‘the highest standard possible and transparent’ (NICE, 2022). According to NICE’s methods guidance for HTA, there are three main requirements for evidence generated in CEA: (i) the included population, comparators and outcomes should be relevant to the evaluation, (ii) the study should use appropriate methods to minimise bias, and (iii) the evidence should be generated in a transparent and reproducible way (NICE, 2022).

Such evidence might come from CEA with various study designs, which raise different issues. Many CEA use Randomised Controlled Trials (RCTs) for the assessment of short-term effectiveness, and the random assignment of the comparators can balance all confounding factors (Willan and Briggs, 2006). However, the availability of RCT evidence in many decision contexts and disease areas is limited. For the evaluation of non-pharmaceutical technologies (e.g., devices or diagnostics), or the introduction of changes to health services, health policy of public health interventions, there are major challenges in generating RCT evidence, which may not be mandated for introduction of the intervention into the health system. By contrast new pharmaceutical agents often require RCT evidence on safety and efficacy prior to marketing authorisation (Skivington et al., 2021). However, RCTs designed for the purposes of marketing authorisation, may have strict eligibility criteria and a short duration of follow-up, which limits the relevance of the evidence generated for the purposes of HTA. Hence, CEA are almost always required to use observational (non-randomised) data within a general modelling framework, in particular for the estimation of parameters pertaining to longer term utilities and resource use, which are unlikely to be appropriately estimated within the limited follow-up period of most RCTs (Briggs et al., 2006).

For the assessment of new health technologies, NICE and other HTA agencies are moving away from the exclusive reliance on RCT evidence for the assessment of relative effectiveness, and towards further use of Real-World Data (RWD) in their evaluations. In this thesis, the definition of RWD is data ‘generated through routine clinical practice and without any intervention by the researcher’ (Garrison et al., 2007;

Makady et al., 2017b). This definition of RWD encompasses data from a broad variety of sources that is routinely collected for purposes not limited to research, such as the reimbursement of health service providers.[4] Under this definition, both registry data and administrative data including electronic health records (EHR) are considered RWD. In this thesis the main interest is in the form of RWD that is collated for administrative purposes, which can also be referred to as 'routine data' (Garrison et al., 2007). Following precedents in the literature, I will refer to real-world evidence as 'evidence generated in observational studies through the analysis of RWD' (Garrison et al., 2007; US FDA, 2018).

The latest methods guidance from NICE highlights the general use of RWD as a priority methods research area, stating "we aim to harness the principles of data science to further our knowledge, using big data and real-world data (RWD) for the benefit of the wider health and social care system". The recently-published NICE real-world evidence framework described potential uses of RWD, good research practices and recommendations for improving transparency and trustworthiness of real-world evidence (NICE, 2022). However, some challenges remain, in particular around the design and conduct of CEA in settings with unmeasured confounding and heterogeneity (see definitions in section 1.2). This thesis seeks to address this gap, by improving methods for assessing comparative effectiveness and cost-effectiveness that use RWD.

The next section describes the potential uses of RWD in HTA processes, and in particular, for the purposes of generating evidence on comparative effectiveness and cost-effectiveness. I outline some common methods for generating this type evidence using RWD, and identify important gaps in these literatures.

1.2 The use of Real-World Data in Cost-effectiveness Analyses

The increased availability of RWD has created opportunities for informing HTA processes in settings where RCT evidence is unavailable or inadequate. A recent review of the policies of six HTA agencies on use of real-world evidence showed that there is

[4] In this thesis, I use the terms 'real-world data' and 'routine data' interchangeability to emphasise that focus is on data is collected in routine care. While some forms of RWD such as genomic data or patient reported outcomes are collected outside of routine clinical care, for practical purposes, I don't make a distinction between these two definitions.

substantial variation in the extent to which international HTA agencies rely on real-world evidence for assessment treatment effectiveness, and for estimating other requisite parameters for CEA (Makady et al., 2017a). NICE's real-world evidence framework describe how it has expanded its criteria for evidence, and how RWD might be used to inform guidance. Table 1.1 provides a non-exhaustive list of current uses of RWD by NICE.

Table 1.1. Uses of real-world data (RWD) for Health Technology Assessment (HTA) and examples from previous guidance from the National Institute of Health and Care Excellence (NICE) - adapted from NICE's real-world evidence framework

Uses of RWD in HTA	Example of NICE guidance
Describing the decision context	HST15 (NICE, 2021)
Informing parameters in economic models	NG115 (NICE, 2019)
Supplementing network meta-analyses in settings where the network is 'incomplete'	TA383 (NICE, 2016)
Generating comparative- or cost-effectiveness evidence within uncontrolled studies such as single-arm trials	HST14 (NICE, 2021a)
Generating comparative- or cost-effectiveness evidence exclusively from RWD sources.	TA524 (NICE, 2018)

Current uses of RWD in technology evaluations include gaining an understanding of the particular decision context (row 1 in Table 1.1), as well as informing parameters in economic models (row 2 in Table 1.1). Uses of RWD in generating comparative effectiveness or cost-effectiveness evidence are not limited to one design. Applications include settings in which RWD is used to generate external comparison groups for single arm trials, settings in which relative costs and outcomes need to be evaluated over a time-horizon beyond the RCT follow-up period, or where the RCT population does not represent the target population of interest (rows 3 and 4 in Table 1.1). A final setting in which RWD could be particularly important is if, in the complete absence of relevant RCT evidence, individual-level RWD is used directly to estimate comparative effectiveness and cost-effectiveness (row 5 in Table 1.1).

This thesis is concerned with this last use of RWD, that is, generating comparative-effectiveness and cost-effectiveness evidence using only RWD, and in the specific setting where individual participant data is available for all the comparison groups of interest. This type of evidence is particularly useful in those settings in which RCT evidence is often unavailable, including public health interventions, non-pharmacological interventions such as surgical procedures, or pharmacological interventions such as treatments for orphan diseases or complex interventions.

Observational studies, including those that use RWD, have the potential to meet NICE's requirement (i) for evidence generated in CEA that 'the included population, comparators and outcomes should be relevant to the evaluation' (NICE, 2022a). However, the reliance on non-randomised designs for assessing comparative effectiveness raises concerns around the risk of bias (requirement ii), and about whether the findings have been generated in a transparent and reproducible way (requirement iii). These concerns were acknowledged in NICE's real-world evidence framework, where the risk of bias from confounding, and the lack of trust in evidence from RWD studies were described as two of the three main barriers to a wider adoption of real-world evidence in HTA decision-making (NICE, 2022). This thesis is concerned with improving methods to help address the first two barriers to the better use and broader adoption of real-world evidence in HTA decision-making. The third barrier alluded to by NICE pertaining to concerns around the quality of the data, is not directly addressed by the thesis, but I will consider issues pertaining to the data quality within the case studies included in the thesis.

Tackling the 'trust barrier' requires that studies improve the transparency and traceability of study design choices. Previous good practice recommendations include using reporting checklists as well as facilitating access to data to help evidence users judge the quality of the evidence. Several checklists and quality assessment tools have been developed for CEA, but they are insufficient for evaluating study design choices in studies using RWD. Most checklists have focussed on related but different issues pertaining to RCTs and decision models (Drummond et al., 2005; Husereau et al., 2013; Philips et al., 2006). While some checklists have been developed that relate to RWD (Kreif et al., 2013), further consideration of broader issues of study design, analysis and interpretation are required. Other recommendations such as pre-registering the study protocol and health economics analysis plan (HEAP), and using structured reporting templates, should also be adopted as they are well-known good research practices, but they are unlikely to help evaluate the risk of confounding in a given study. More recently, NICE's 2022 real-world evidence framework recommended that the notions of the 'target trial' framework should be adopted in observational studies.

The target trial framework was developed in the epidemiological literature as a tool to help minimise the risk of bias and design flaws in observational studies (Hernán and Robins, 2016). The target trial is a hypothetical trial that the researcher would design to evaluate the research question. The target trial framework requires researchers to specify crucial standpoints to the analysis of RWD, such as stating the study eligibility criteria, and defining the comparator groups. By encouraging researchers to apply the design principles of RCTs, the target trial framework can help reduce the risk of bias from using inadequate study designs to assess comparative effectiveness (Dickerman et al., 2019; Petito et al., 2020). The application of the notions of the target trial can also allow evidence-users judge the design choices made in the study by formally evaluating how closely the study design emulates that of an analogous RCT (Dahabreh et al., 2020; García-Albéniz et al., 2017; Lodi et al., 2019). Gomes et al., (2022) describes ways the target trial framework could be applied to alternative study designs and uses of RWD in HTA, but there is a lack of guidance and exemplar applications in CEA that use RWD to evaluate treatment effectiveness and cost-effectiveness.

While the target trial framework offers some general principles for the design of observational studies, on its own, this framework is insufficient to mitigate the inevitable concerns about confounding in observational studies (second barrier). Confounding arises when baseline covariates associated with the outcome are not balanced between treatment strategies (Hernán et al., 2002). In studies using RWD, treatment strategies are not randomised and have a high risk of unmeasured confounding (or residual confounding), that is, confounding due to unmeasured baseline characteristics. A key purpose of the study design and methods of analysis for assessments of comparative effectiveness and cost-effectiveness that use RWD is to minimise the risk of confounding.

An advantage of RWD is that it can target the population of interest for decision-making purposes, that is, patients who present for the health care interventions in question in routine clinical practice. However, the inclusion within the RWD of a broader population than those who would enrol in an RCT, may raise concerns about 'treatment effect heterogeneity' in addition to those of confounding (Bell et al., 2016; Sarri et al., 2022). In particular, if drivers of treatment effect heterogeneity are also associated with the choice of treatment strategy, and the outcome of interest, then the study needs appropriate methods to account for the confounding effect of those variables. In health care, both measured and unmeasured prognostic factors such as the patient's age or the stage of the disease may be expected to influence treatment selection and also explain the individual's response to treatment. In this thesis, I describe 'overt (treatment effect) heterogeneity' as heterogeneity that is according to measured characteristics within the RWD. I refer to 'essential (treatment effect) heterogeneity' as heterogeneity according to unmeasured prognostic variables.[5]

This thesis is concerned with evaluations of comparative effectiveness and cost-effectiveness in presence of essential heterogeneity. This is a common phenomenon in health care research, and raises important concerns for observational studies evaluating treatment effects, as treatment choice is often according to patient characteristics such as the patient's capacity to benefit from either treatment strategy which is unlikely to be measured within RWD. However, although this problem is common, methods to tackle both confounding and heterogeneity due to unmeasured characteristics have not been well-developed, in the setting of comparative effectiveness and cost-effectiveness studies.

1.3 Methods for evaluating treatment effects in Cost-Effectiveness Analyses

In the general causal inference, biostatistics and econometrics literature, numerous methods have been developed to address the risk of confounding inherent in observational studies (Hernán and Robins, 2020; Pearl, 2000). In CEA, some progress has been made in the transfer of methods from these general literatures to address specific issues raised in this context such as the joint distribution of endpoints (Nixon and Thompson, 2005; Polsky and Basu, 2012; Sekhon and Grieve, 2012). Broadly, these methods can be grouped into methods that assume 'no unmeasured confounding' (this is often referred to as the 'unconfoundedness' assumption), and those that do not rule out the possibility of unmeasured confounding. Methods in the first group such as regression adjustment have been widely adopted in CEAs (see, for example, Kreif et al. (2012); Nixon and Thompson (2005); Willan et al., (2004)). These methods are generally appropriate for estimating policy-relevant treatment effect parameters such as the Average Treatment Effect (ATE) in settings where the treatment assignment mechanism is well-understood, and it is plausible to assume that the important confounding factors are measured in the data.

In settings where the adjustment for observed prognostic factors is unlikely to provide sufficient protection against bias due to confounding, using Instrumental Variable (IV) methods might be advisable. IV methods can provide reliable estimates of treatment

[5] This terminology is used for consistency with the existing health econometrics literature.

effects even in presence of unmeasured confounding provided some requisite assumptions about the validity and relevance of the instrument hold (Baiocchi et al., 2014; Brookhart et al., 2015; Rassen et al., 2009). The properties of IV methods under these assumptions have been discussed in the econometrics literature (Baiocchi et al., 2014; Brookhart et al., 2015), and they have been extensively adopted in the applied economics literature. However, their use is still relatively uncommon in health care, including in CEA (see, for example, Prentice et al., 2014; Saramago et al., 2020).

The appropriateness of traditional IV methods for estimating treatment effects of decision-making relevance such as ATE or CATEs, largely depends on the form of treatment effect heterogeneity that is present (Angrist et al., 1993; Angrist and Fernández-Val, 2011). For instance, in presence of essential heterogeneity, the Two-Stage Least Squares (2SLS) estimator can provide consistent estimates of the Local Average Treatment Effect (LATE), which is the average treatment effect among an unidentifiable subgroup of individuals in the population, but not necessarily the ATE for the population (Basu et al., 2007). Instead, Local instrumental variable (LIV) methods can provide robust estimates of comparative effectiveness that apply to policy-relevant populations in presence of essential heterogeneity, provided some assumptions hold (Heckman and Vytlacil, 2001). In Chapter 2, I provide an overview of the IV methodology, including the identification assumptions, with particular attention to LIV, which is the primary focus of this thesis.

1.4 Case study: the ESORT study

The methodological contributions of this thesis were motivated and informed by the 'Emergency Surgery OR noT' (ESORT) study. This was a study funded by the National Institute for Health and Care Research (NIHR) that sought to evaluate the outcomes, costs and cost-effectiveness of emergency surgery (ES) for patients with common acute conditions (ESORT Study Group, 2020). This section provides: an overview of the ESORT study, focussing on the aspects that are relevant to the thesis; a description of my contribution to the ESORT study; and a brief explanation of the ESORT study to help define some of the specific objectives of the thesis.

1.4.1 Overview of the ESORT study

The ESORT study was a retrospective cohort study that used routine data from hospital episode statistics (HES) for emergency admissions to NHS hospitals in England to evaluate the effectiveness and cost-effectiveness of ES compared to alternative non-emergency surgery (NES) strategies, such as delayed surgery or antibiotic therapy, for patients with five common acute conditions: acute appendicitis, diverticular disease, acute gallstone disease, abdominal wall hernia, and intestinal obstruction.

For these conditions, RCT evidence on the benefits, risks and costs associated with the provision of ES is scarce (Azhar et al., 2021; Flum et al., 2020; Javanmard-Emamghissi et al., 2021; Thornell et al., 2016). Observational studies have failed to address the major concern of unmeasured confounding (Koumarelas et al., 2014; Saverio et al., 2014). Clinical advisors to the ESORT project raised the concern that the decision as to whether patients have ES or the NES alternative is associated with baseline factors that are prognostic of outcomes such as all-cause mortality at 90 days. Hence unless these differences between the comparison groups are measured, and allowed for, the study would provide biased estimates of the effectiveness of ES, due to confounding by indication. These baseline factors may also modify the relative effectiveness of ES, and include some that are measured within the data, such the patient's age, which can lead to overt heterogeneity. However, other baseline factors that are not measured in HES, such as the severity of the disease, which can modify the effectiveness of ES, i.e., essential heterogeneity is a major potential concern.

The lack of evidence to inform clinical guidelines pertaining to the choice of strategy for patients presenting as emergency admissions with these acute conditions has resulted in wide variation in rates of ES across NHS hospitals in England (Abercrombie, 2017). The ESORT study sought to address this gap in the literature by exploiting this natural variation in use of ES. The ESORT study built on a precedent study that used an IV design to address confounding in evaluating the effects of ES versus non-operative strategies in the United States (Keele et al., 2018). The precedent study found that there were no substantial differences in clinical outcomes following ES versus NES strategies at the aggregated population level, but that for some pre-specified subgroups of patients, NES could lead to better outcomes (Keele et al., 2018). However, while this precedent study was useful in supporting the IV design taken in the ESORT study it did not consider the cost-effectiveness of ES nor did it consider an LIV approach to address essential heterogeneity.

1.4.2 Study data

This thesis used HES admitted patient care data on emergency admissions to 175 NHS hospitals in England (Herbert et al., 2017). Data was provided to the ESORT study team under a data-sharing agreement with NHS Digital. The data comprised emergency admissions and any subsequent readmissions of adult patients between 1 April 2009 and 30 June 2020. Mortality data was obtained from linkage of Office for National Statistics (ONS) death records with HES. The data included rich clinical and sociodemographic information, including the patient's age, gender and index of multiple deprivation (IMD). Information on medical interventions and surgical procedures was also available, as well as administrative information such as dates of surgical procedures and ultimate hospital discharge. Health-related quality of life and unit cost data were derived from the literature (see Chapter 4 for further details).

1.4.3 Contribution of the candidate to the ESORT study

Prior to the start of the thesis, the ESORT study had not started, and as aspects of the study design and analysis plans had not been specified, I was able to contribute in the following areas, pertaining to the thesis. First, I was able to develop and apply the target trial framework to the ESORT study. This required me to identify issues raised in adapting the general target trial framework to the HTA setting, to work with the project team to devise solutions, and to draft the resulting paper (see Chapter 3). Second, the application of the LIV framework to the ESORT study, required me to consider carefully the requisite assumptions pertaining to IV in general, and LIV in particular. I conducted the LIV analyses, alongside one of my supervisors, led the interpretation of the CEA results, and drafted the accompanying paper (Chapter 4). Third, motivated by the initial findings of the ESORT study, I led an extensive simulation study looking at the properties of the LIV approach in settings with different forms of heterogeneity, and with scenarios motivated by the ESORT study. I interpreted the results and drafted the resulting paper (Chapter 5). For each of the three empirical papers for the thesis I include a statement which clearly delineates my own contribution from those of other ESORT team members including my PhD supervisors (see chapters 3, 4 and 5).

1.5 Aims and objectives of the thesis

The main aim of this thesis is to help address some of the gaps in methods for CEA that use routine data. The broad research question that this thesis sought to answer is: "*Can Local Instrumental Variables methods inform CEAs with reliable estimates of relative effectiveness and cost-effectiveness in the presence of unmeasured* *confounding and treatment effect heterogeneity?"* To be able to answer this question, I defined the following three research objectives of this thesis:

1. *Critically examine the application of the principles of the target trial framework to the HTA context, identify the main challenges, and provide recommendations to address them.*
   This was a novel application of the target trial framework within the HTA context. In Chapter 3, I provide an illustration of how this methodology can help to minimise concerns about confounding and design flaws in CEAs. I describe some of the main challenges for studies using RWD, and recommendations to address them. This study, in its paper format, is currently being considered for publication in *Value in Health (March 2023).*

2. *Evaluate and implement an LIV approach for addressing unmeasured confounding and heterogeneity in CEA.*
   Chapter 4 includes an application of LIV in a CEA using routine data on emergency surgery admissions to NHS hospitals in England. The study formally evaluates the identification assumptions for LIV, and contrasts this methodology with alternative regression adjustment and IV approaches. Chapter 4 was published in *Medical Decision Making (May 2022).*

3. *Evaluate the performance of different IV approaches in terms of bias and statistical efficiency according to alternative levels of IV strength, sample sizes and forms of heterogeneity in a simulation study.*
   To achieve this objective, a study was conducted using Monte Carlo simulation methods to measure the bias and efficiency implications for LIV of different levels of instrument strength, sample sizes, and forms of treatment effect heterogeneity. The results of this study can be found in Chapter 5 of the thesis. This study is currently being considered for publication in *Health Economics (March 2023).*

1.6 Overall contribution of the thesis

All three research objectives have been met through three research papers. The three papers have been submitted to journals, and have either been published (research paper 2 was published in *Medical Decision Making* (Moler-Zapata et al., 2022) or will be considered for publication (research papers 1 and 3 are currently being considered for publication in the journals *Value in Health* and *Health Economics*, respectively).

Research paper 1 describes the main challenges for comparative effectiveness and cost-effectiveness studies that apply the target trial framework using RWD. These challenges relate to different aspects of the target trial's design, including the definition of the eligibility criteria, treatment strategies and time zero. The paper also considers the major risk of confounding, which is one of the main concerns for CEA, and comparative effectiveness studies more generally that use RWD. I argue that carefully evaluating the risk of these issues in the study design, and applying the recommendations outlined in research paper 1, will not only help the study minimise the risk of confounding, but will help evidence users to judge whether the resulting evidence is adequate to inform the research question. These recommended practices could help improve the trustworthiness of real-world evidence, and facilitate its timely adoption in HTA and policy-making. The main recommendations describe how to plug gaps in the RWD using expert clinical judgement, for example in emulating the trial's treatment eligibility criteria to ensure comparable populations across treatment groups; and how to use novel IV methods for estimation and inference on treatment effect parameters of decision-making relevance.

Research paper 2 makes two important contributions to existing methods for CEA. First, the paper illustrates how RWD can be used to identify continuous instruments for use in real-world applications. LIV methods using a continuous IV constructed using routine data are used to evaluate policy-relevant treatment effects. Second, the study contrasts alternative IV methods that target different treatment effect parameters, and make different assumptions about confounding and heterogeneity, and I evaluate them in the context of the ESORT study.

Research paper 3 addresses the gap in the guidance for applied LIV studies in terms of IV strength requirements in conjunction with different available sample sizes. The study builds on insights from research paper 2 to evaluate how LIV performs in terms of bias and statistical efficiency (measured by the root mean squared error, rMSE) in estimating the ATE and CATE parameters. I consider different scenarios defined by the strength of the instrument, the sample size and the form of treatment effect heterogeneity. The main contribution of the study is in demonstrating that the LIV approach provides estimates for ATE and CATE with lower levels of bias and RMSE, irrespective of the sample size or IV strength, compared to the 2SLS method. The study also finds that, in general, with smaller sample sizes, both methods require stronger instruments to ensure low levels of bias.

1.7 Structure of the thesis

The remaining chapters of the thesis are as follows. Chapters 3 to 5 comprise the three research papers, each with a preamble within which I define my specific contribution.

Chapter 3 (research paper 1) describes the main challenges in applying the notions of target trial framework to CEA that use routine data to inform HTA decision-making using working examples from the ESORT study. I offer recommendations for future studies looking to apply the target trial framework in evaluations of the effectiveness and cost-effectiveness of health interventions. Chapter 4 (research paper 2) builds on the preceding chapter in using the ESORT study to provide an exemplar application of the LIV methodology within a CEA that studies heterogeneity in outcomes and costs across patient characteristics. I describe the key methodological aspects of the LIV methodology, including the target estimand, and the identification assumptions underlying the methodology. I also demonstrate how RWD can be used to test some of these assumptions. Chapter 5 (research paper 3) draws motivation from the ESORT study to define a simulation study in which the reliability of LIV is evaluated according to how the method performs across settings with different sample sizes, levels of IV strength and forms of treatment effect heterogeneity. I contrast the performance of the LIV approach against that of the method of 2SLS in the simulation study, but also in cohorts derived using data from the ESORT study. Chapter 6 provides an overview of the main findings and contributions of the thesis. The chapter acknowledges the limitations of the thesis, and identifies the main areas for future research. This chapter concludes by highlighting the implications of the findings of the thesis for applied researchers and policy makers. Appendices are available at the end of the thesis, and references at the end of each chapter.

References

Abercrombie J (2017) *Getting it Right First Time (GiRFT) report. General Surgery.* Available at: http://gettingitrightfirsttime.co.uk/national-general-surgery-report-published-2/.

Angrist J, Imbens G and Rubin D (1993) Identification of causal effects using instrumental variables. *Journal of the American Statistical Association* 91(434): 444–455.

Angrist JD and Fernández-Val I (2011) ExtrapoLATE-ing: External validity and overidentification in the LATE framework. *Advances in Economics and Econometrics: Tenth World Congress Volume 3, Econometrics*: 401–434. DOI: 10.1017/CBO9781139060035.012.

Azhar N, Johanssen A, Sundström T, et al. (2021) Laparoscopic Lavage vs Primary Resection for Acute Perforated Diverticulitis: Long-term Outcomes From the Scandinavian Diverticulitis (SCANDIV) Randomized Clinical Trial. *JAMA Surgery* 156(2): 121–128. DOI: 10.1001/jamasurg.2020.5618.

Baiocchi M, Cheng J and Small DS (2014) Instrumental variable methods for causal inference. *Statistics in Medicine* 33(13): 2297–2340. DOI: 10.1002/sim.6128.

Basu A (2014) Estimating person-centered treatment (PeT) effects using instrumental variables: an application to evaluating prostate cancer treatments. *JOURNAL OF APPLIED ECONOMETRICS* 29: 671–691. DOI: 10.1002/jae.

Basu A, Heckman JJ, Navarro-Lozano S, et al. (2007) Use of instrumental variables in the presence of heterogeneity and self-selection: an application to treatments of breast cancer patients. Health Economics 16(2007): 1133–1157. DOI: 10.1002/hec.1291.

Bell H, Wailoo AJ, Hernandez M, et al. (2016) *The use of real world data for the estimation of treatment effects in NICE decision making: Report by the Decision Support Unit.*

Bjorklund A and Moffitt R (1987) The Estimation of Wage Gains and Welfare Gains in Self-Selection Models. *The Review of Economics and Statistics* 69(1): 42. DOI: 10.2307/1937899.

Briggs A, Claxton K and Sculpher M (2006) *Decision Modelling for Health Economic Evaluation.* Handbooks in health economic evaluation series. Oxford: Oxford University Press.

Brookhart MA, Rassen JA and Schneeweiss S (2015) *Instrumental variable methods in comparative safety and effectiveness research. Effective Health Care Research Report No. 22.* Rockville, MD. Available at: http://effectivehealthcare.ahrq.gov/reports/final.cfm.

Claxton K, Palmer S, Sculpher M, et al. (2010) *Appropriate perspectives for health care decisions, working Paper No 054cherp.* Centre for Health Economics, University of York.

Culyer AJ (2010) *The Dictionary of Health Economics.* London: Edward Elgar Publishing.

Dahabreh IJ, Robins JM and Hernán MA (2020) Benchmarking observational methods by comparing randomized trials and their emulations. *Epidemiology* 31(5): 614–619. DOI: 10.1097/EDE.0000000000001231.

Dakin H, Devlin N, Feng Y, et al. (2015) The influence of cost-effectiveness and other factors of NICE decisions. *Health Economics* 24: 1256–1271. DOI: 10.1002/hec.

Dickerman BA, García-Albéniz X, Logan RW, et al. (2019) Avoidable flaws in observational analyses: an application to statins and cancer. *Nature Medicine* 25(10). Springer US: 1601–1606. DOI: 10.1038/s41591-019-0597-x.

Drummond MF, Sculpher MJ, Torrance GW, et al. (2005) *Methods for the Economic Evaluation of Health Care Programmes*. Oxford medical publications. Oxford University Press. Available at: https://books.google.es/books?id=CxWzQgAACAAJ.

Eckermann S and Pekarsky B (2014) Can the real opportunity cost stand up: Displaced services, the straw man outside the room. *PharmacoEconomics* 32(4): 319–325. DOI: 10.1007/s40273-014-0140-3.

ESORT Study Group (2020) Emergency Surgery Or NoT (ESORT) study. Available at: https://www.lshtm.ac.uk/media/38711.

Flum DR, Davidson GH, Monsell SE, et al. (2020) A Randomized Trial Comparing Antibiotics with Appendectomy for Appendicitis. *New England Journal of Medicine* 383(20): 1907–1919. DOI: 10.1056/nejmoa2014320.

García-Albéniz X, Hsu J, Hernán MA, et al. (2017) The value of explicitly emulating a target trial when using real world evidence: an application to colorectal cancer screening HHS Public Access. *European Journal of Epidemiology* 32(6): 495–500. DOI: 10.1007/s10654-017-0287-2.

Garrison LP, Neumann PJ, Erickson P, et al. (2007) Using real-world data for coverage and payment decisions: The ISPOR real-world data Task Force report. *Value in Health* 10(5). International Society for Pharmacoeconomics and Outcomes Research (ISPOR): 326–335. DOI: 10.1111/j.1524-4733.2007.00186.x.

Gomes M, Latimer N, Soares M, et al. (2022) Target trial emulation for transparent and robust estimation of treatment effects for health technology assessment using real-world data: opportunities and challenges. *PharmacoEconomics*. Springer: 1–10.

Heckman JJ and Vytlacil EJ (1999) Local instrumental variables and latent variable models for identifying and bounding treatment effects. *Proceedings of the National Academy of Sciences of the United States of America* 96: 4730–4734. DOI: 10.1073/pnas.96.8.4730.

Heckman JJ and Vytlacil EJ (2001) Policy-Relevant Treatment Effects. *American Economic Review* 91(2): 107–111. DOI: 10.1257/aer.91.2.107.

Herbert A, Wijlaars L, Zylbersztejn A, et al. (2017) Data Resource Profile: Hospital Episode Statistics Admitted Patient Care (HES APC). *International Journal of Epidemiology* 46(4): 1093-1093i. DOI: 10.1093/ije/dyx015.

Hernán MA and Robins JM (2006) Instruments for causal inference: An epidemiologist's dream? *Epidemiology* 17(4): 360–372. DOI: 10.1097/01.ede.0000222409.00878.37.

Hernán MA and Robins JM (2016) Using Big Data to Emulate a Target Trial When a Randomized Trial Is Not Available. *American Journal of Epidemiology* 183(8). Oxford University Press: 758–764. DOI: 10.1093/aje/kwv254.

Hernán MA and Robins JM (2020) *Causal Inference: What If.* Boca Raton: Chapman & Hall/CRC.

Hernán MA, Hernández-Díaz S, Werler MM, et al. (2002) Causal knowledge as a prerequisite for confounding evaluation: an application to birth defects epidemiology. *American journal of epidemiology* 155(2). Oxford University Press: 176–184.

Husereau D, Drummond M, Petrou S, et al. (2013) Consolidated Health Economic Evaluation Reporting Standards (CHEERS) statement. *European Journal of Health Economics* 14: 367–372. DOI: 10.1007/s10198-013-0471-6.

Javanmard-Emamghissi H, Hollyman M, Boyd-Carson H, et al. (2021) Antibiotics as first-line alternative to appendicectomy in adult appendicitis: 90-day follow-up from a prospective, multicentre cohort study. *British Journal of Surgery*: 1–9. DOI: 10.1093/bjs/znab287.

Keane M and Neal T (2021) *A Practical Guide to Weak Instruments. UNSW Economics Working Paper No. 2021-05d.*

Keele L, Sharoky CE, Sellers MM, et al. (2018) An instrumental variables design for the effect of emergency general surgery. *Epidemiologic Methods* 7(1). Walter de Gruyter GmbH. DOI: 10.1515/em-2017-0012.

Koumarelas K, Theodoropoulos GE, Spyropoulos BG, et al. (2014) A prospective longitudinal evaluation and affecting factors of health related quality of life after appendectomy. *International Journal of Surgery* 12(8). Elsevier Ltd: 848–857. DOI: 10.1016/j.ijsu.2014.06.015.

Kreif N, Grieve R, Radice R, et al. (2012) Methods for estimating subgroup effects in cost-effectiveness analyses that use observational data. *Medical Decision Making* 32(6): 750–763. DOI: 10.1177/0272989X12448929.

Kreif N, Grieve R and Sadique MZ (2013) Statistical methods for cost-effectiveness analyses that use observational data: a critical appraisal tool and review of current practice. *Health Economics* 22: 486–500. DOI: 10.1002/hec.

Lee D, McCrary J, Moreira MJ, et al. (2021) *Valid T-Ratio Inference for IV. National Bureau of Economic Research Working Paper Series (No. w29124).* DOI: 10.2139/ssrn.3901588.

Lodi S, Phillips A, Lundgren J, et al. (2019) Effect Estimates in Randomized Trials and Observational Studies: Comparing Apples With Apples. *American Journal of Epidemiology* 188(8). Oxford Academic: 1569–1577. DOI: 10.1093/AJE/KWZ100.

Makady A, Ham R, de Boer A, et al. (2017a) Policies for Use of Real-World Data in Health Technology Assessment (HTA): A Comparative Study of Six HTA Agencies. *Value in Health* 20(4). Elsevier Inc.: 520–532. DOI: 10.1016/j.jval.2016.12.003.

Makady A, de Boer A, Hillege H, et al. (2017b) What Is Real-World Data? A Review of Definitions Based on Literature and Stakeholder Interviews. *Value in Health* 20(7). Elsevier Inc.: 858–865. DOI: 10.1016/j.jval.2017.03.008.

Moler-Zapata S, Grieve R, Lugo-Palacios D, et al. (2022) Local instrumental variable methods to address confounding and heterogeneity when using electronic health records: an application to emergency surgery. *Medical Decision Making* 0(0). SAGE Publications Inc STM: 0272989X221100799. DOI: 10.1177/0272989X221100799.

National Institute for Health and Care Excellence (2016) *TA383: TNF-alpha inhibitors for ankylosing spondylitis and non-radiographic axial spondyloarthritis*. London (UK). Available at: ww.nice.org.uk/guidance/ TA383.

National Institute for Health and Care Excellence (2018) *TA524: Brentuximab vedotin for treating CD30-positive Hodgkin lymphoma*. London (UK).

National Institute for Health and Care Excellence (2019) *NG115: Chronic obstructive pulmonary disease in over 16s: diagnosis and management*. London (UK). Available at: https://www.nice.org.uk/guidance/ng115.

National Institute for Health and Care Excellence (2021a) *HST14: Metreleptin for treating lipodystrophy*. London.

National Institute for Health and Care Excellence (2021b) *HST15: Onasemnogene abeparvovec for treating spinal muscular atrophy*. London (UK). Available at: https://www.nice.org.uk/guidance/hst15.

National Institute for Health and Care Excellence (2022a) *NICE health technology evaluations: the manual*. London.

National Institute for Health and Care Excellence (2022b) *NICE real-world evidence framework*. London (UK). Available at: www.nice.org.uk/corporate/ecd9.

Neyman J (1990) On the application of probability theory to agricultural experiments. *Statistical Science* 5: 463–480.

Nixon RM and Thompson SG (2005) Methods for incorporating covariate adjustment, subgroup analysis and between-centre differences into cost-effectiveness evaluations. *Health Economics* 14: 1217–1229. DOI: 10.1002/hec.1008.

Ochalek J, Revill P and Drummond M (2020) Allocating Scarce Resources—Tools for Priority Setting. In: *Global Health Economics: Shaping Health Policy in Low-and Middle-Income Countries*. World Scientific, pp. 53–73.

Pearl J (2000) *Models, Reasoning and Inference*. New York: Cambridge University Press.

Petito LC, García-Albéniz X, Logan RW, et al. (2020) Estimates of Overall Survival in Patients With Cancer Receiving Different Treatment Regimens: Emulating Hypothetical Target Trials in the Surveillance, Epidemiology, and End Results (SEER)-Medicare Linked Database. *JAMA network open* 3(3): e200452. DOI: 10.1001/jamanetworkopen.2020.0452.

Philips Z, Ginnelly L, Sculpher M, et al. (2006) Review of guidelines for good practice in decision-analytic modelling in health technology assessment. *Health Technology Assessment* 8(36). DOI: 10.3310/hta8360.

Polsky D and Basu A (2012) *Chapter 46: Selection Bias in Observational Data.* (AM Jonesed.). he Elgar C. Edward Elgar Publishing.

Rassen JA, Brookhart MA, Glynn RJ, et al. (2009) Instrumental variables I: instrumental variables exploit natural variation in nonexperimental data to estimate causal relationships. *Journal of Clinical Epidemiology* 62(12): 1226–1232. DOI: 10.1016/j.jclinepi.2008.12.005.Instrumental.

Rubin DB (1974) Estimating causal effects of treatment in randomized and nonrandomized studies. *Journal of Educational Psychology* 66(5): 688–701. Available at: http://www.fsb.muohio.edu/lij14/420_paper_Rubin74.pdf.

Rudrapatna VA and Butte AJ (2020) Opportunities and challenges in using real-world data for health care. *The Journal of Clinical Investigation* 130(2). Am Soc Clin Investig: 565–574.

Sarri G, Bennett D, Debray T, et al. (2022) ISPE-Endorsed Guidance in Using Electronic Health Records for Comparative Effectiveness Research in COVID-19: Opportunities and Trade-Offs. *Clinical Pharmacology & Therapeutics*. Wiley Online Library.

Saverio S Di, Sibilio AM, Giorgini E, et al. (2014) The NOTA Study (Non Operative Treatment for Acute Appendicitis): Prospective Study on the Efficacy and Safety of Antibiotics (Amoxicillin and Clavulanic Acid) for Treating Patients With Right Lower Quadrant Abdominal Pain and Long-Term Follow-up of Conser. *Annals of Surgery* 260(1): 109–117. DOI: 10.1136/bmjopen-2010-000006.

Sekhon JS and Grieve R (2012) A matching method for improving covariate balance in cost-effectiveness analyses. *Health Economics* (21): 695–713. DOI: 10.1002/hec.1748.

Skivington K, Matthews L, Simpson SA, et al. (2021) Framework for the development and evaluation of complex interventions: Gap analysis, workshop and consultation-informed update. *Health Technology Assessment* 25(57). DOI: 10.3310/HTA25570.

Sorenson C, Drummond M, Kristensen FB, et al. (2008) *How can the impact of health technology assessments be enhanced?* Tallinn: World Health Organization. Regional Office for Europe.

Thornell A, Angenete E, Bisgaard T, et al. (2016) Laparoscopic Lavage for Perforated Diverticulitis With Purulent Peritonitis. *Annals of Internal Medicine* 164(3): 137–145. DOI: 10.7326/M15-1210.

US Food and Drug Administration (2018) *Framework for FDA's real-world evidence program.* Available at: https://www.fda.gov/media/120060/download.

Willan AR and Briggs AH (2006) *Statistical Analysis of Cost-Effectiveness Data.* John Wiley & Sons.

Willan AR, Briggs AH and Hoch JS (2004) Regression methods for covariate adjustment and subgroup analysis for non-censored cost-effectiveness data. Health Economics 13(5): 461–475. DOI: 10.1002/hec.843.

World Health Organization (2000) The World Health Report 2000: Health Systems: Improving Performance. Geneva, Switzerland

Chapter 2. Methods

2.1 Overview of Instrumental Variable methods

The main goal of CEAs is to produce reliable estimates of relative effectiveness, costs and cost-effectiveness for the overall target population of interest, the ATE. It is also important to understand how heterogeneity can inform the stratification or personalisation of treatment choices. Incorporating this type of evidence into decision-making processes can result in improved patient outcomes and gains in efficiency (Basu and Meltzer, 2007; Espinoza et al., 2014). In settings where treatment effects are suspected to be modified by observed patient characteries, CEAs are increasingly reporting Conditional Average Treatment Effects (CATEs) –i.e., treatment effects for groups of patients defined by values a particular observed prognostic factor–, alongside estimates of an overall ATE for the population. To be able to inform reliable estimates of ATE and CATEs, studies need to use appropriate techniques for addressing unmeasured confounding.

IV designs can be used to address the concern of unmeasured confounding in CEAs using RWD (see Baiocchi et al., 2014; Brookhart et al., 2006 and Martens et al., 2006 for reviews). A good instrument should meet the following criteria: (i) it is associated with treatment receipt (relevance condition), (ii) it affects the outcome only through its association with the treatment (exclusion restriction condition), (iii) it is not associated with unmeasured confounders (exchangeability condition), and (iv) the direction of the association with the treatment must be the same, irrespective of the level of the IV (monotonicity). Provided the existence of a valid and sufficiently strong instrument, IV methods can provide reliable estimates of treatment effect parameters (Angrist et al., 1993; Baiocchi et al., 2014). However, the likely risk of bias in settings with essential heterogeneity is problematic, and evaluating treatment effect parameters often involves trading off different assumptions by different methods.

In this chapter, I introduce notation and describe the main IV approaches considered in the thesis. This intends to be a brief overview of the key standpoints of the IV methodology. For further details about how these methods were applied in the thesis, I refer the interested readers to Chapters 3, 4 and 5.

2.2 Notation and structural models

Following Heckman and Vytlacil (1999, 2001), I consider a model for the outcome based on the Neyman-Rubin potential outcomes framework and define model a latent variable discrete choice model for selection into treatment (Neyman, 1990; Rubin, 1974). I let treatment be a binary variable, D_z, take values 1 and 0, depending on whether the individual receives the treatment. $Y_0 = \mu_0(X_0, X_U, \vartheta)$ and $Y_1 = \mu_1(X_0, X_U, \vartheta)$ represent the potential outcomes under treatment and control, where X_0 is a vector of observed confounders, X_U is a vector of unmeasured confounders, and ϑ captures any remaining unobserved variation. $\Delta = Y_1 - Y_0$ is the individual treatment effect.

I consider the following model for treatment assignment,

$$D_z{}^* = \gamma(X_0, Z) \geq U_D \text{ and,}$$

$$D_z = 1 \text{ if } D_z{}^* \geq 0 \text{ and } D_z = 0\text{, otherwise}$$

Where $D_z{}^*$ is the 'latent' propensity for treatment, Z is a vector of instruments, and U_D reflects 'distaste' for treatment, and captures the effect of X_U and other variables that discourage treatment assignment. Following Heckman and Vytlacil (1999), and without loss of generality, we can express this model in terms of probabilities as, $D_z{}^* = P(X_0, Z) - V$, where $P(X_0, Z) = F_{U_D|x_O,z}[\gamma(X_O, Z)]$ is the propensity for treatment based on the observed characteristics, and where $V = F_{U_D}[X_{U_D}|X_O = x_O, Z = z]$ with $V \perp (Z, X_O)$ reflects the degree to which unobserved variables discourage treatment, and is uniformly distributed between 0 and 1.

2.3 Conventional IV methods

Imbens and Angrist (1994) and Angrist et al. (1993) introduced the Local Average Treatment Effect (LATE) parameter. The LATE can be defined as $\Delta^{LATE}(x_o, z, z') = E[Y_1 - Y_0|X_O = x_o, D_z < D_{z'}]$, and under the assumptions listed above, it can be identified by the IV estimand:

$$\frac{E[Y|X_O = x_o,\ Z = z'] - E[Y|X_O = x_o,\ Z = z]}{E[D|X_O = x_o,\ Z = z'] - E[D|X_O = x_o,\ Z = z]}$$

The LATE is the average effect for the subgroup of individuals in the population whose treatment status changes as the IV shifts from z to z'. The subpopulation for whom $D_z < D_{z\prime}$ holds are often referred to as 'compliers' (Baiocchi et al., 2014).

The 2SLS (Wald) estimator is the most widely used method for estimating linear models. It is implemented in two stages. In the first stage (or reduced form), D_z is regressed on X_O and Z to obtain estimates of $E[D_Z|X_O, Z]$. In the second stage, Y_D is regressed on X_O and $\hat{E}[D_Z|X_O, Z]$ to obtain an unbiased estimate of $E[Y_D|D_z, X_O, Z]$. When the instrument is continuous, pairwise combinations of z and z' will produce different LATEs. However, as discussed in section 1.3, for 2SLS to inform policy-relevant treatment effects such as the ATE or CATEs, it is required that there is no essential heterogeneity (Heckman et al., 2006).

Alternatively, the Two-Stage Residual Inclusion (2SRI) method can retrieve estimates of the ATE even in presence of essential heterogeneity, similar to the control function approach (Terza et al., 2008; Wooldridge, 2010). The main difference compared to 2SLS, is that this estimator uses the residuals from a first-stage regression for treatment assignment when fitting the model for Y_D, which is regressed on the D_z, X_O and the residuals, which might be included in different forms (Basu et al., 2018). This approach is analogous to 2SLS when both stages are linear, but has been mostly applied in non-linear settings (Basu et al., 2018). It is unclear whether 2SRI offers additional benefits in terms of bias reductions in estimates of ATE or LATE parameters compared to 2SLS. As discussed in Basu et al., (2018), while logit or probit models might offer a better fit to real-world data, 2SRI estimates could be biased if the functional form of the residuals is misspecified.

Novel LIV methods constitute an attractive alternative for estimating treatment effect parameters of decision-making relevance when a continuous or multi-valued IV is available.

2.4 Local Instrumental Variables methods

Heckman and Vytlacil (1999, 2001, 2005) showed that LIV methods can identify effects for "marginal" patients, that is, patients who are in equipoise with respect to the treatment assignment decision, provided a valid, continuous instrument is available. These individuals are in equipoise because the propensity for treatment, given their observed levels of covariates and IV, just balance with a normalized version of the unmeasured confounders (V) discouraging treatment, such that a small (marginal) change in the IV is sufficient to nudge them into the treatment group.

Then, by contrasting individuals with marginally different values of the IV, but who are otherwise identical in measured and unmeasured covariates, the Marginal Treatment Effect (MTE) can be identified (Bjorklund and Moffit, 1987).

2.4.1 Marginal Treatment Effects

The MTE can be defined as,

$$\Delta^{MTE}(x_O, v) = E(\Delta | X_O = x_O, V = v)$$

The MTE is the most nuanced treatment effect parameter. MTEs can be seen as building blocks that can be used to compute the ATE or LATE. When the MTE is constant in U_D –i.e. patients do not act upon the unobserved confounders–, then $\Delta^{MTE} = \Delta^{ATE} = \Delta^{LATE}$. Under essential heterogeneity, the different treatment effects can be computed as weighted averages of Δ^{MTE}.

Under standard IV assumptions, streams of MTEs can be estimated as (Heckman and Vytlacil, 2001),

$$\Delta^{MTE}(x_O, p) = \frac{\partial E(Y_1 - Y_0 | X_O = x_O, Z = z)}{\partial p}$$

LIV recovers MTEs for all the values in the support of the distribution of $P(Z)$ conditional on $X_O = x_O$.

2.4.2 Person-centered Treatment effects

Basu (2014) extended the LIV framework to consider personalised treatment effects known as Person-centered Treatment (PeT) effects. PeT effects can be derived from MTEs by using information on the observed patient characteristics, and the likely distribution of unobserved characteristics given the patient's observed treatment status. The underlying insight is that for each individual patient, some levels of the normalized unobserved confounder would be inconsistent with the observed treatment decision for that individual, given their observed characteristics and the level of the IV (Basu, 2014). For patients in the treatment group ($D = 1$), the propensity to choose treatment based on X_O and Z outweighs the propensity to choose the comparator strategy based on V, i.e. $P(z, x_O) > v$, whereas the opposite is true for patients in the comparator strategy ($D = 0$). MTEs that imply a lower level of unobserved confounding can thus be 'ruled out', narrowing the set of MTEs which could plausibly represent the individual's effect. The person-centered treatment (PeT) effect for an individual is obtained by aggregating the remaining MTEs.

Hence,

$$\Delta^{PeT}(x_O, p, D) = E(Y_1 - Y_0 | X_O = x_O, P(z, x_O) > V) \text{ for individuals with } D = 1$$

$$\Delta^{PeT}(x_O, p, D) = E(Y_1 - Y_0 | X_O = x_O, P(z, x_O) < V) \text{ for individuals with } D = 0$$

PeT effect averages MTEs with the same level of X_O and Z over those values of the unobserved confounders that are compatible with that patient's treatment assignment. For individuals with $D = 1$, PeT effects can be derived as,

$$E(Y_1 - Y_0 | X_O = x_O, P(z, x_O) > V) = P(z)^{-1} \int_0^{P(z)} MTE(x_O, v) dv$$

All treatment effect parameters, including CATEs, can be derived by taking averages of PeT effects. This is therefore a well-suited approach exploring treatment effect heterogeneity, but requires that a valid, continuous or multi-valued IV is available (Basu, 2014).

2.4.3 Estimation of MTEs and PeT effects

In this thesis, I follow the approach described in Basu (2014, 2015) to estimate MTEs and PeT effects using the LIV methodology. Briefly, D_z is regressed on Z and X_O, using appropriate methods for binary outcomes, to obtain an estimate of the propensity for treatment, or propensity score, $\hat{p}(x_O, z)$. At this stage, an F statistic[6] test should be performed to evaluate the strength of the IV. Next, Y_D is regressed on X_O and a function of $\hat{p}$ including interactions with X_O. The approach outlined in Basu (2014) involves differentiating the outcome model $g(Y_D)$ by $\hat{p}(x_O, z)$. Next, PeT effects for each individual can be obtained by performing numerical integration, with MTE ($\partial \hat{g}(Y_D)/\partial \hat{p}\,(x_O, z)$) evaluated by replacing $\hat{p}$ using 1,000 random draws of $u \sim unif(\min(\hat{p}(x_O, z)), \max(\hat{p}(x_O, z)))$. Then, $D^* = \Phi^{-1}\{\hat{p}(x_O, z)\} + \Phi^{-1}(1 - u)$ can be computed. PeT effects can then be computed by averaging $\partial \hat{g}(Y_D)/\partial \hat{p}\,(x_O, z)$ over values of u for which $\hat{p}(x_O, z) > v$ if $D = 1$; or over values of $\hat{p}(x_O, z) < v$ if $D = 0$. Finally, averaging PeT effects over all of the observations provides an estimate of the

[6] The F statistic can be computed as $F = N^*(\hat{\pi}/\hat{\sigma}_\pi)$, with $\hat{\pi}$ and $\hat{\sigma}_\pi$ as the estimated value of the coefficient of Z in the first stage and its associated standard error.

ATE for the population, and over strata of X_O gives the CATE for the subpopulation of interest. Standard errors can be computed using bootstrap methods (Basu, 2015). The Stata developed "`petiv`" command was used in this thesis to estimate PeT effects.

2.5 Problems with weak IVs

The advent of RWD has created opportunities for adopting LIV methods in comparative effectiveness and cost-effectiveness studies. One important barrier for a wider adoption of LIV methods is that it might have poor estimation and inference properties if the IV is only weakly associated with treatment assignment (Staiger and Stock, 1997, Andrews et al., 2019).

The implications this might have for practice have not yet been formally evaluated. Some recently published papers in the weak identification literature have demonstrated the shortcomings of relying exclusively on the 'rule of thumb' that the F-statistic in the first stage needs to be above the threshold value of 10 in the case of binary IVs. These studies have shown that even when IVs are considered 'strong' by conventional standards, 2SLS can have low power (Keane and Neal, 2021), as well as size distortions in t-tests (Lee et al., 2021). These findings suggest that even when IVs meet conventional thresholds for 'strength', 2SLS might be unreliable. However, no studies have formally evaluated the IV strength required for LIV methods to perform well, nor whether requirements change according to the sample size, or the form of treatment effect heterogeneity that is present.

The next chapter considers the target trial framework in the context of the ESORT study using LIV. In doing so, this work constitutes a novel application of the target trial paper in a IV study.

References

Andrews I, Stock JH and Sun L (2019) Weak Instruments in Instrumental Variables Regression: Theory and Practice. *Annual Review of Economics* 11: 727–753. DOI: 10.1146/annurev-economics-080218-025643.

Angrist J, Imbens G and Rubin D (1993) Identification of causal effects using instrumental variables. *Journal of the American Statistical Association* 91(434): 444–455.

Baiocchi M, Cheng J and Small DS (2014) Instrumental variable methods for causal inference. *Statistics in Medicine* 33(13): 2297–2340. DOI: 10.1002/sim.6128.

Basu A (2014) Estimating person-centered treatment (PeT) effects using instrumental variables: an application to evaluating prostate cancer treatments. *Journal of applied econometrics* 29: 671–691. DOI: 10.1002/jae.

Basu A (2015) Person-centered treatment (PeT) effects: Individualized treatment effects using instrumental variables. *The Stata Journal* 15(2): 397–410.

Basu A and Meltzer D (2007) Value of information on preference heterogeneity and individualized care. *Medical Decision Making* 27(2): 112–127. DOI: 10.1177/0272989X06297393.

Basu A, Coe NB and Chapman CG (2018) 2SLS versus 2SRI: Appropriate methods for rare outcomes and/or rare exposures. *Health Economics* 27(6): 937–955. DOI: 10.1002/hec.3647.

Bjorklund A and Moffitt R (1983) *Estimation of Wage Gains and Welfare Gains from Self-Selection Models. IUI Working Paper, No. 105,*. Stockholm.

Brookhart MA, Wang PS, Solomon DH, Schneeweiss S. Evaluating short-term drug effects using a physician-specific prescribing preference as an instrumental variable. Epidemiology. 2006;17(3):268-275. doi:10.1097/01.ede.0000193606.58671.c5

Espinoza MA, Manca A, Claxton K, et al. (2014) The Value of Heterogeneity for Cost-Effectiveness Subgroup Analysis: Conceptual Framework and Application. *Medical Decision Making* 34(8): 951–964. DOI: 10.1177/0272989X14538705.

Heckman JJ and Vytlacil EJ (1999) Local instrumental variables and latent variable models for identifying and bounding treatment effects. *Proceedings of the National Academy of Sciences of the United States of America* 96: 4730–4734. DOI: 10.1073/pnas.96.8.4730.

Heckman JJ and Vytlacil EJ (2001) Policy-Relevant Treatment Effects. *American Economic Review* 91(2): 107–111. DOI: 10.1257/aer.91.2.107.

Heckman JJ and Vytlacil E (2005) Structural equations, treatment effects and econometric policy evaluation. *Econometrica* 73(3): 669–738. DOI: 10.1111/j.1468-0262.2005.00594.x.

Heckman JJ, Urzua S and Vytlacil E (2006) *Understanding instrumental variables in models with essential heterogeneity. NBER Working Paper No. 12574.* Cambridge, MA. DOI: 10.1162/rest.88.3.389.

Imbens GW and Angrist JD (1994) Identification and Estimation of Local Average Treatment Effects. *Econometrica* 62(2). JSTOR: 467. DOI: 10.2307/2951620.

Keane M and Neal T (2021) *A Practical Guide to Weak Instruments. UNSW Economics Working Paper No. 2021-05d.*

Lee D, McCrary J, Moreira MJ, et al. (2021) *Valid T-Ratio Inference for IV. National Bureau of Economic Research Working Paper Series (No. w29124).* DOI: 10.2139/ssrn.3901588.

Martens EP, Pestman WR, De Boer A, Belitser S V., Klungel OH. Instrumental variables: Application and limitations. Epidemiology. 2006;17(3):260-267. doi:10.1097/01.ede.0000215160.88317.cb

Neyman J (1990) On the application of probability theory to agricultural experiments. *Statistical Science* 5: 463–480.

Prentice JC, Conlin PR, Gellad WF, et al. (2014) Capitalizing on prescribing pattern variation to compare medications for type 2 diabetes. *Value in Health* 17(8). Elsevier: 854–862.

Rubin DB (1974) Estimating causal effects of treatment in randomized and nonrandomized studies. Journal of Educational Psychology 66(5): 688–701. Available at: http://www.fsb.muohio.edu/lij14/420_paper_Rubin74.pdf.

Saramago P, Claxton K, Welton NJ, et al. (2020) Bayesian econometric modelling of observational data for cost-effectiveness analysis: establishing the value of negative pressure wound therapy in the healing of open surgical wounds. *Journal of the Royal Statistical Society. Series A: Statistics in Society* 183(4): 1575–1593. DOI: 10.1111/rssa.12596.

Staiger D and Stock JH (1997) Instrumental Variables Regression with Weak Instruments. *Econometrica* 65(3): 557. DOI: 10.2307/2171753.

Terza J V., Basu A and Rathouz PJ (2008) Two-stage residual inclusion estimation: Addressing endogeneity in health econometric modeling. *Journal of Health Economics* 27(3). NIH Public Access: 531–543. DOI: 10.1016/j.jhealeco.2007.09.009.

Wooldridge, JM. (2010) Econometric Analysis of Cross Section and Panel Data. The MIT Press, JSTOR, http://www.jstor.org/stable/j.ctt5hhcfr. Accessed 17 Mar. 2023.

Chapter 3. Emulating Target Trials with Real World Data to inform Health Technology Assessment: findings and lessons from an application to emergency surgery

3.1 Preamble to research paper 1

The target trial framework was developed to help improve the design of comparative effectiveness studies using observational data, by emulating the design principles of RCTs with respect to, for example, the eligibility criteria or the comparator strategies (Hernán and Robins 2016). Since then, there has been a rapid increase in the number of studies, mainly in the biostatistics and pharmaco-epidemiological literature, that have used the methods described in Hernán and Robins (2016) and, Hernán et al. (2016). Previous studies, including early efforts from the RCT DUPLICATE initiative have sought to apply the target trial framework in the design and analysis of observational studies to replicate RCTs. These studies have found that while RCTs can be replicated using RWD, further research is needed to better understand the circumstances or contexts in which this real-world evidence will align with RCT evidence (Franklin et al. 2021; Danaei et al. 2018).

Applications of the target trial framework to RWD for the purposes of informing HTA decision-making are uncommon. Recently, Gomes et al. (2022) described the potential uses of the framework for informing HTA processes. However, this paper did not offer an exemplar application, or give recommendations for future studies on how to address the challenges that might arise in CEA of health interventions using RWD.

Research paper 1 paper aims to fill this gap by evaluating the challenges raised for CEA using individual-participant RWD, when no relevant RCT evidence is available. I draw from the main findings of the paper to offer recommendations for how to address these challenges in future studies. My role in this paper included reviewing the relevant literatures, developing and applying the target trial framework in the study, and conducting the analyses, guided by my supervisor, RG. I led the interpretation of the results. I wrote the draft version of the manuscript and incorporated comments from co-authors, SON, AH, RS and RG, into the manuscript. The analysis received ethical approval from the LSHTM Ethics Committee (ID:21776).

References

Danaei G, García Rodríguez LA, Cantero OF, et al. (2018) Electronic medical records can be used to emulate target trials of sustained treatment strategies. *Journal of Clinical Epidemiology* 96. Elsevier USA: 12–22. DOI: 10.1016/j.jclinepi.2017.11.021.

Franklin JM, Patorno E, Desai RJ, et al. (2021) Emulating Randomized Clinical Trials with Nonrandomized Real-World Evidence Studies: First Results from the RCT DUPLICATE Initiative. *Circulation*: 1002–1013. DOI: 10.1161/CIRCULATIONAHA.120.051718.

Gomes M, Latimer N, Soares M, et al. (2022) Target trial emulation for transparent and robust estimation of treatment effects for health technology assessment using real-world data: opportunities and challenges. *PharmacoEconomics*. Springer: 1–10.

Hernán MA and Robins JM (2016) Using Big Data to Emulate a Target Trial When a Randomized Trial Is Not Available. *American Journal of Epidemiology* 183(8). Oxford University Press: 758–764. DOI: 10.1093/aje/kwv254.

Hernán MA, Sauer BC, Hernández-Díaz S, et al. (2016) Specifying a target trial prevents immortal time bias and other self-inflicted injuries in observational analyses. *Journal of Clinical Epidemiology* 79: 70–75. DOI: 10.1016/j.jclinepi.2016.04.014.

LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE

London School of Hygiene & Tropical Medicine
Keppel Street, London WC1E 7HT

T: +44 (0)20 7299 4646
F: +44 (0)20 7299 4656
www.lshtm.ac.uk

RESEARCH PAPER COVER SHEET

Please note that a cover sheet must be completed for each research paper included within a thesis.

SECTION A – Student Details

Student ID Number	1903225	**Title**	Ms
First Name(s)	Silvia		
Surname/Family Name	Moler Zapata		
Thesis Title	METHODS TO ADDRESS CONFOUNDING AND HETEROGENEITY IN COST-EFFECTIVENESS ANALYSES USING REAL-WORLD DATA		
Primary Supervisor	Prof Richard Grieve		

If the Research Paper has previously been published please complete Section B, if not please move to Section C.

SECTION B – Paper already published

Where was the work published?			
When was the work published?			
If the work was published prior to registration for your research degree, give a brief rationale for its inclusion			
Have you retained the copyright for the work?*	Choose an item.	Was the work subject to academic peer review?	Choose an item.

*If yes, please attach evidence of retention. If no, or if the work is being included in its published format, please attach evidence of permission from the copyright holder (publisher or other author) to include this work.

SECTION C – Prepared for publication, but not yet published

Where is the work intended to be published?	Value in Heath
Please list the paper's authors in the intended authorship order:	Silvia Moler-Zapata, Andrew Hutchings, Stephen O'Neill, Richard J. Silverwood, Richard Grieve

Stage of publication	**Not yet submitted**

SECTION D – Multi-authored work

For multi-authored work, give full details of your role in the research included in the paper and in the preparation of the paper. (Attach a further sheet if necessary)	My role in this paper included reviewing the relevant literatures, developing and applying the target trial framework in the study, and conducting the analyses, guided by RG. I led the interpretation of the results. I wrote the draft version of the manuscript and incorporated comments from co-authors, SON, AH, RS and RG, into the manuscript.

SECTION E

Student Signature	Silvia Moler Zapata
Date	23 Sep. 22

Supervisor Signature	Richard Grieve
Date	23 Sep. 22

3.2 Research paper 1: Emulating Target Trials with Real World Data to inform Health Technology Assessment: findings and lessons from an application to emergency surgery

Authors

Silvia Moler-Zapata[1], Andrew Hutchings[1], Stephen O'Neill[1], Richard J. Silverwood[2], Richard Grieve[1]

Affiliations

[1] Department of Health Services Research and Policy, London School of Hygiene & Tropical Medicine, London, UK

[2] Centre for Longitudinal Studies, UCL Social Research Institute, University College London, London, UK

Abstract

Objective: International Health Technology Assessment (HTA) agencies continue to advocate for the use of real-world data (RWD) for informing decision-making in health care. There is potential for the 'target trial' framework to encourage further uptake of this type of evidence by helping to alleviate concerns about bias and design flaws in these studies through the application of the design principles of randomised controlled trials. So far, its adoption in HTA has been modest, arguably due to the lack of guidance and exemplar implementations in this particular setting.

Methods: We apply the two-stage (definition and emulation) target trial emulation approach in a study assessing the cost-effectiveness of emergency surgery for two acute gastrointestinal conditions (acute appendicitis and acute gallstone disease). We use hospital episodes statistics (HES) data for emergency hospital admissions with acute these conditions to 175 acute hospitals in England from 2010 to 2019. We highlight and describe the main challenges in applying the target trial framework in studies using RWD, and discuss how these were addressed in this particular application.

Results: Our study identifies four main challenges for RWD studies applying the target trial framework. These are: (i) defining the study population, (ii) defining the treatment strategies, (iii) establishing time zero (baseline), and (iv) adjusting for unmeasured confounding. We exemplify how these challenges were addressed within the 'Emergency Surgery OR noT' (ESORT) study and, drawing on these findings, we outline a series of recommendations for how they can be addressed more widely when using the target trial framework alongside RWD.

Conclusion: Studies using the target trial framework are likely to face similar issues to those that that arose in the ESORT study, and discussed here. The recommendations outlined in this study could help future studies, and should be considered complementary to design tools developed for economic evaluations to inform HTA, as well as those developed for informing choices about the adequacy of alternate statistical approaches for estimating treatment effects.

Keywords: real-world data, target trial framework, health technology assessment, comparative effectiveness, emergency surgery.

3.2.1 Introduction

Health Technology Assessment (HTA) agencies require robust effectiveness and cost-effectiveness evidence to support decision-making in health care. Studies using Real World Data (RWD) such as disease registry data or electronic health records (EHR) can help build an evidence base, given their ability to include patients from large, heterogenous populations, and offer results for interventions of decision-making relevance and broad ranges of outcomes (Garrison et al., 2007; Makady et al., 2017a). However, the risk of bias from confounding and other design flaws in these studies constitute a major barrier to a more widespread adoption of real-world evidence in HTA decision-making (Bell et al., 2016; Faria et al., 2015).

Good research practices recommendations by HTA agencies like the UK's National Institute for Health and Care Excellence (NICE) include the use of checklists and other quality assessment tools, or the reporting health economic analysis plans, but these offer limited guidance on how to address fundamental issues pertaining to study design of studies using RWD (Husereau et al., 2013; Thorn et al., 2016). Recently, NICE's latest manual of methods and processes for technology evaluation formally recognised the importance of RWD in informing decision-making and emphasised the need for studies that consider how the principles of the 'target trial' framework could be applied to HTA (Hernán et al., 2016; Hernán and Robins, 2016).

The target trial framework can help mitigate concerns about the study design in observational (non-experimental) studies by applying the design principles of Randomised Controlled Trials (RCTs) (Dickerman et al., 2019; Hernán et al., 2016). This approach requires the definition of a (hypothetical) pragmatic trial protocol, which is then emulated using observational data. The target trial framework can help to better identify and minimise the risk of bias in the study, and make methodological assumptions and design choices transparent for evidence users. In settings with high quality observational data analyses, emulating the target trial principles has been found to help replicate the results of published RCTs (Caniglia et al., 2018; Franklin et al., 2021; Petito et al., 2020). More recently, Gomes et al., (2022) described the potential uses of the target trial framework in HTA, but did not actually use the methods in an application. In general, there is a lack of guidance on the apply the notions of the target trial framework in the HTA context, which raises major challenges, in particular around the interrelated issues of defining from RWD the study population, time zero (baseline) and the intervention and comparators.

The aim of this paper is to critically examine the application of the target trial framework principles to the HTA context when assessing the effectiveness of health interventions from RWD. We draw on a case-study, the ‘Emergency Surgery OR noT’ (ESORT) study, to describe common challenges in applying the target trial framework to assess comparative effectiveness from routine data and offer a series of recommendations for future studies (ESORT Study Group, 2020a). Unlike previous publications of the ESORT study (Hutchings et al., 2022; Moler-Zapata et al., 2022), here we define and emulate the key elements of the target trial protocol, in evaluating the cost-effectiveness of emergency surgery (ES) for patients admitted to hospital with acute gastrointestinal conditions (section 3.2.2), and report the results of the CEA (section 3.2.3). In section 3.2.4, we draw on these findings to offer general recommendations for future studies.

3.2.2 Methods

3.2.2.1 Overview

The ESORT study exemplifies the challenges that arise for HTA when there is little evidence from RCTs to inform routine clinical practice. In this setting there were few published RCTs and economic evaluations that evaluated ES versus alternative non-emergency surgery (NES) strategies for common acute conditions (ESORT Study Group, 2020a). The ESORT study helped address this gap in the literature by using information on 2010-19 hospital admissions from the Hospital Episode Statistics (HES) database, linked to Office for National Statistics (ONS) mortality data, to assess the cost-effectiveness of ES for five acute gastrointestinal conditions, including acute appendicitis and acute gallstone disease, which are the two conditions with the highest prevalence, and the focus of this paper. The evaluation of costs and outcomes was from a hospital perspective, over a one-year time horizon, and applied the key principles of the target trial framework as described in the following sections (Moler-Zapata et al., 2022).

3.2.2.2 Target population for the decision problem

The application of the target trial framework to the HTA context, requires that eligibility criteria for the study population are defined to represent the target population of interest and that it only includes those subgroups for whom there is equipoise about the choice of intervention versus comparator strategies (i.e., clinical uncertainty about which treatment alternative is the best option for them). Only patients who would in practice be eligible to receive either intervention, even if one is more likely than the other, should be included.

In the ESORT study, these considerations informed the choice of inclusion and exclusion criteria (Table 3.1). Some inclusion criteria, such as the patient's age and the requirement to be assessed by a surgeon, were intended to ensure equipoise and were emulated directly from the HES data. The clinical panel was asked to identify subgroups of patients who, according to unobserved as well as observed prognostic characteristics, would not be eligible to receive ES or NES. For example, by specifying an inclusion criterion that the patient must be at some point under the care of a consultant surgeon, the study deliberately excluded patients whose prognosis was so 'severe' or 'mild' according to unobserved, as well as observed, characteristics, that they would not be considered for ES. Including patients for whom there is no equipoise could result violations of the positivity assumption if there is insufficient or no variability in treatment within strata of confounders (Petersen et al., 2012). For other criteria, such as the reason for admission, the information from the routine data and the available evidence were insufficient to define which patient subgroups to include (Table 3.1) lists those criteria that could not be directly emulated using RWD). Specifically, while there was information on the patients' diagnosis according to ICD-10 codes, it was unclear which of the subcategories of ICD-10 corresponded to patient subgroups that would be eligible for ES in routine practice, and for whom there was equipoise between the comparison strategies.

The ESORT study addressed the challenge of defining those elements of the target trial protocol that could not be specified from the routine data, by convening two panels of 12 clinicians with relevant expertise that followed a modified Delphi process (see ESORT Study Group (2020b) for details). The panellists were required to judge which inclusion and exclusion criteria were appropriate, given the requirement for equipoise between the comparison groups, and to define the interventions of interest (see next section). The consensus of the panel required at least nine from 12 responses in favour of the inclusion of the category, and five (appendicitis) and three (gallstone disease) ICD sub-categories were designated for inclusion (see Appendix B.1) for full list). The panel's consensus also designated that for patients with acute appendicitis those with ICD-10 codes corresponding to appendiceal cancer and pregnancy should be excluded due to lack of equipoise, but for patients with gallstone disease none of the subcategories should be excluded.

Table 3.1. Protocol of the target trial of emergency surgery (ES) versus non-emergency surgery (NES) for acute appendicitis and acute gallstone disease

	Description of Target Trial of ES	**How was the protocol element emulated in the ESORT?**
Eligibility criteria	Inclusion criteria:	Inclusion criteria:
	– Patient was at least 18 years old at admission.	– Emulated directly from HES data.
	– Emergency admission, via emergency department or primary care.	– Emulated directly from HES data.
	– The condition was the reason for admission into hospital.	– Expert panel defined diagnostic (ICD-10) codes with equipoise between comparator strategies. *
	– The diagnosis was confirmed by consultant surgeon.	– Emulation directly from HES data.
	Exclusion criteria:	Exclusion criteria:
	– According to clinical condition-specific exclusion criteria.	– Expert panel designated exclusion criteria with (ICD-10) codes. †
	– Emergency admission for the condition in the previous year.	– Emulated directly from HES data.
	– Surgery for the condition within the previous 90 days.	– Emulated directly from HES data (using definitions of treatment strategies below).
	– Patient transferred between hospitals before surgical assessment.	– Emulated directly from HES data.
		Additional criteria according to data availability:
		– Patient was admitted to an ineligible hospital for ESORT. ‡
		– Admission lacked information on admission or discharge status or date.
Treatment strategies	– ES defined as urgent, expedited or immediate surgery for the condition (NCEPOD, 2004).	– Expert panel defined the two criteria for ES: (i) the procedure constituted ‘surgery for the condition’ according to selected OPCS codes, § (ii) to be considered ‘emergency’, the panel designated a time window of seven days from the date of assessment (see below).

	Description of Target Trial of ES	How was the protocol element emulated in the ESORT?
	– NES: (i) medical management with no surgery for the condition; (ii) surgery that did not meet the criteria for ES, either because not relevant procedure, or after the seven-day time window, possibly preceded by medical management.	– Emulation assumed patient assigned NES if they did not meet ES criteria
Time zero and follow-up period	– Time zero is analogous to the time of randomisation, and is when all the eligibility criteria are met, the assignment to ES or NES occurs, and follow-up starts.	– Emulation assumed time zero was the start date of the first finished consultant episode for the first admission, in which the specialty code was general surgery, colorectal surgery or upper-gastrointestinal surgery.
	– Follow-up ends at the earliest of one year, death, or end of study period.	– Emulation censored patients at the date of death, if that was within one year from day zero. Complete follow-up data were available for all patients.
Treatment assignment	– Individuals are randomly assigned to a strategy at baseline.	– Treatments groups were assumed to be balanced after adjustment for differences in measured and unmeasured prognostic factors in the statistical analysis.
Outcomes	– Life years at 1 year from randomisation.	– Emulated directly from HES data (linked to ONS death data).
	– QALYs at 1 year from randomisation.	– Emulation required adjusting life years using published age- and gender-adjusted HRQoL scores from similar populations.
	– Total costs at 1 year from randomisation.	– Emulation required calculating resource use for categories considered to be main drivers of total costs (length of stay, including critical care; operative and diagnostic procedures and readmissions up to one year) and valuing resource use data using relevant estimates of unit costs taken from national unit cost databases.
	– Net monetary benefit at 1 year from randomisation.	– Emulated combining cost and QALY data.

	Description of Target Trial of ES	How was the protocol element emulated in the ESORT?
Causal contrast of interest	– ITT effect (effect of assignment of patients to interventions at baseline) – PP effect (effect of complying with the trial protocol)	– ITT effect could not be emulated since information on the initial treatment assignment was not available from HES. – Emulation of the per-protocol effect required taking differences between the treatment groups in estimated total costs, life years, QALYs and net monetary benefits at one year.
Analysis plan	– ITT analysis and PP analysis with adjustment for baseline prognostic factors.	– Emulation of the PP analysis required using a LIV approach to mitigate the risk of confounding due to unmeasured prognostic factors associated with ES receipt. The IV was the hospital's tendency to operate. Models were adjusted for a wide range of case-mix measures (age, gender, frailty level, comorbidity profile, ethnicity, index of multiple deprivation), fixed effects for each financial year and proxies of quality of acute care (rates of emergency admission and mortality for each hospital and acute condition in 2009-10, and in the year prior to the admission).
	– Subgroup analyses by baseline age, sex, frailty and number of comorbidities.	– Emulated directly from HES data.

*See Appendix B.2 for full list ICD10 codes for the two conditions. †ICD-10 codes for acute appendicitis: Pregnancy (O00-O9A; Z00-Z99) and appendiceal cancer (C00-D49). ICD-10 codes for acute gallstone disease: none. ‡ Of all eligible acute general hospitals with at least 200 emergency general surgery admissions per year, those that ceased activity in five years prior to 31 December 2019 were excluded. § See Appendix B.2 for full list of procedure codes included in definition of ES the two conditions (ESORT Study Group, 2020b). FCE: Finished Consultant Episode, HES: Hospital Episode Statistics, HRQoL: Health-related Quality of Life, ICD: International Classification of Diseases, ITT: Intention-to-treat, LIV: Local Instrumental Variables, ONS: Office of National Statistics, OPCS: Office of Population Censuses and Surveys, PP: Per-protocol, QALY: Quality-adjusted Life Years.

3.2.2.3 Definition of treatment strategies

The main challenge in defining the treatment strategies from the RWD is ensuring that these represent how the intervention is used in routine clinical practice. In the ESORT study, the treatment strategies under assessment were complex, combining different surgical and non-surgical procedures. ES involves operative management that is immediate, urgent or expedited (NCEPOD, 2004). To operationalise the ES definition, the expert panel were asked to consider which of the Office for Population Censuses and Surveys (OPCS) procedure codes listed within the HES data met the definition of ES, and to define the appropriate time window. The panel's consensus was that 21 (appendicitis) and 45 (gallstones) procedure codes (see Table 3.1 and Appendix B.2), respectively, met the definition for ES, and that for both conditions the time window for ES should be within seven days of assessment (baseline/time zero, see below).

The definition of the comparator strategy should consider whether the information in the RWD is sufficient to ensure the comparator strategy is defined in enough detail to evaluate the causal contrast of interest (Hernán, 2004; Hernán and Taubman, 2008). In the ESORT study, any patient who didn't receive one of the designated procedures within the 7-day period was assigned to the NES strategy. This definition includes management with antibiotic therapy and either no surgery within the one-year time horizon, or surgery that does not meet the ES criteria (i.e., either an OPCS procedure code not considered to be ES within the designated ES window or an OPCS procedure code considered to be ES but outside the window). The proposed definition of the comparator strategy in ESORT, reflects the variation in the provision NES strategies in routine clinical practice, but also the limited availability of granular information in HES on specific NES treatments (e.g., duration or dosage for antibiotic therapy), which meant that the study could evaluate the cost-effectiveness of ES against not providing NES, but not against specific NES strategies.

3.2.2.4 Definition of time zero and follow-up

The careful definition of the emulated target trial's 'point of randomisation' or 'time zero' can help minimise the risk of bias in the study (Emilsson et al., 2018; Hernán et al., 2016). In an RCT, time zero is defined as the time when eligibility is met, the alternative treatment strategies commence, and the follow-up begins. In RWD studies, it is often impossible to establish temporality from events recorded in the data, and if eligibility and treatment assignment are not aligned with the start of follow-up, then selection bias (if patients are excluded according to events that occurred after the onset of treatment) and immortal time bias (if there is a period of the follow-up over which outcomes of interest cannot occur) can emerge (Lévesque et al., 2010; Maringe et al., 2020). The criteria for time zero are: (i) it does not precede the time when the eligibility criteria are met, (ii) it must be identified for all patients regardless of the assigned treatment arm strategy, (iii) it should minimise the time window used to define treatment initiation to reduce the possibility of immortal time bias.

In ESORT, emulating time zero was not straightforward. The study considered using the date of hospital admission or the date either strategy was initiated, but both were deemed inadequate. For many patients, the date of admission preceded the date diagnosis was confirmed by a surgeon which was an inclusion criterion (violation of i). Also, for the NES comparator, a date of treatment initiation was not available (violation of ii). A third alternative, the date that the patient was first under the care of a consultant surgeon was judged to be the most appropriate definition of time zero. After this initial surgical assessment, patients with these acute conditions would be assigned to either treatment strategy, without delay. Given the study's eligibility criteria, once the patient had the surgical assessment all the eligibility criteria were met. This definition of time zero could still lead to bias, if during the seven-day time window for defining receipt of ES (rather than NES), the risk of the outcomes of interest differed between the comparison groups. For patients with acute appendicitis and acute gallstone disease, this would seem unlikely as patients are at very low risk of adverse outcomes, such as death, over that period (Di Saverio et al., 2020). When assessing ES for other conditions with higher rates of in-hospital mortality, methods like 'cloning, censoring and weighting' could help to reduce the risk of immortal time bias (Hernán and Robins, 2016).

3.2.2.5 Outcomes

The nature of RWD might pose additional challenges for the emulation of the target trial as data on outcome measures are often unavailable or available with insufficient detail. However, through HES, the ESORT study had access to rich resource use data including the total duration of hospital stay (including readmissions), and survival time from HES linked to ONS mortality data, which was used to derive life years. While information on health-related quality of life (HRQoL) following ES and NES for the conditions was not available from HES, it could be obtained from available published studies reporting HRQoL weights for comparable populations. These weights were combined with information on key events (e.g., emergency admissions), and survival time to derive one-year Quality-Adjusted Life Years (QALYs) (Moler-Zapata et al., 2022). The main cost-effectiveness outcome is the incremental net monetary benefit (INB) at one year, using NICE's recommended threshold of £20,000 per QALY (NICE, 2013).

3.2.2.6 Causal contrast

RCTs are typically concerned with estimating the Intention-to-treat effect (ITT), that is, the effect of being assigned to a particular treatment strategy and the per-protocol effect (PP), that is, the effect of receiving the treatment as prescribed in the protocol. In observational studies, where treatment received is observed but treatment assignment is not, a PP analysis is generally favoured. In the ESORT study, the broad protocol definition of both the ES and the NES strategies allowed us to estimate a PP effect. Here, the assumption that patients in either group adhered to their treatment assignment is plausible and consistent with routine practice.

3.2.2.7 Analysis plan

The risk of confounding bias poses a major threat to the validity of observational studies, and alternative methods make different assumptions, which need to be carefully considered (Freemantle et al., 2013). The ESORT study used a Local Instrumental Variable (LIV) approach to mitigate the concerns about unmeasured confounding. Briefly, LIV allows for treatment selection according to measured and unmeasured prognostic factors, and can report consistent estimates of the overall effect for the population (i.e., the Average Treatment Effect, ATE) and subpopulations of interest (i.e., conditional ATEs, CATEs) provided a series of assumptions hold (Moler-Zapata et al., 2022).

The instrument in the ESORT study was the hospital's tendency to operate (TTO), which is a proxy for their preference for ES, calculated from historic data. The assumptions underlying LIV are: (i) TTO only influences the outcome through its effect on treatment assignment (exclusion restriction), (ii) TTO is associated with treatment assignment (relevance assumption), (iii) TTO is independent of unmeasured confounders (exchangeability condition), and (iv) TTO has the same direction of effect on the probability of treatment receipt, irrespective of the level of the IV (monotonicity assumption). These assumptions were judged plausible, given the findings that the IV was sufficiently strong (assumption ii), balanced the observed covariates (iii) and by implication and a priori reasoning also unobserved covariates

(i) and, was unlikely to have a differential effect on the probability of ES receipt at different levels of TTO (iv) (see Appendix B.5).

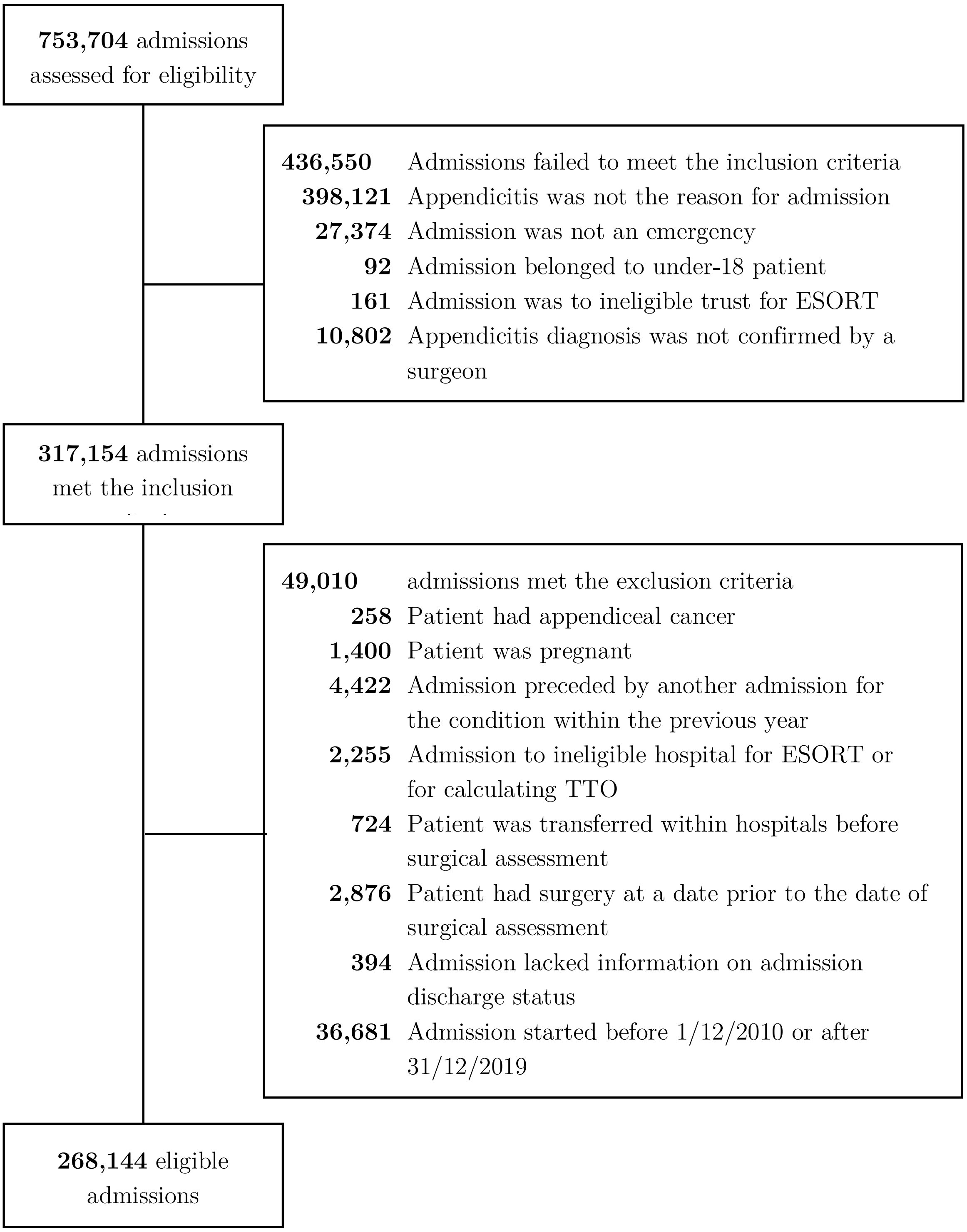
We also conducted analyses which made alternative assumptions as sensitivity analyses. We undertook conventional risk-adjustment (using generalised linear model (GLM) regression) approaches, adjusting for the same baseline measures as in the LIV analysis, but that makes the alternative assumption that all the requisite confounders have been adjusted for. For completeness we also included a naïve comparison, that assumed there were no confounders. For each approach we reported the incremental net monetary benefit (INB) for the overall target population of interest (ATE), and for LIV the INB according to prespecified subgroups of policy relevance (defined by age, sex, frailty level and number of comorbidities.

3.2.3 Results

3.2.3.1 Cohort description

We identified 268,144 patients with acute appendicitis and 240,977 with gallstone disease who met the target trial eligibility (see Figure 3.1). Of these patients, 92% (appendicitis) and 22% (gallstone disease) met the definition of ES, and the baseline characteristics of the comparison groups are given in see Table 3.2. In each cohort, those patients who had ES were on average younger, fitter and with fewer comorbidities (Table 3.2).

Figure 3.1. (a): Flowchart of eligibility for a target trial of emergency surgery versus non-emergency surgery for acute appendicitis, emulation using Hospital Episodes Statistics data



753,704 admissions assessed for eligibility

436,550 Admissions failed to meet the inclusion criteria
398,121 Appendicitis was not the reason for admission
27,374 Admission was not an emergency
92 Admission belonged to under-18 patient
161 Admission was to ineligible trust for ESORT
10,802 Appendicitis diagnosis was not confirmed by a surgeon

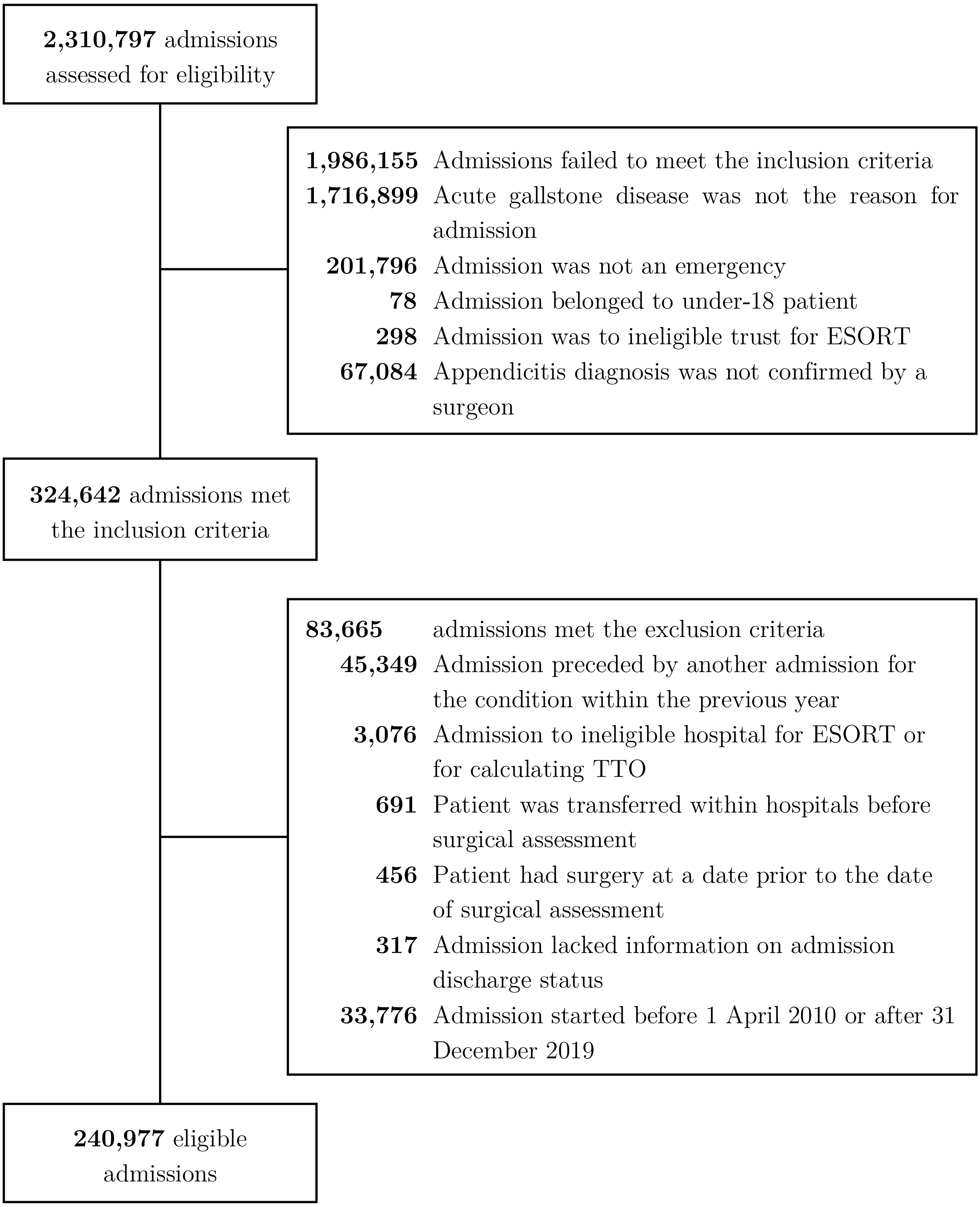
317,154 admissions met the inclusion

49,010 admissions met the exclusion criteria
258 Patient had appendiceal cancer
1,400 Patient was pregnant
4,422 Admission preceded by another admission for the condition within the previous year
2,255 Admission to ineligible hospital for ESORT or for calculating TTO
724 Patient was transferred within hospitals before surgical assessment
2,876 Patient had surgery at a date prior to the date of surgical assessment
394 Admission lacked information on admission discharge status
36,681 Admission started before 1/12/2010 or after 31/12/2019

268,144 eligible admissions

A&E: accident and emergency, GP: general practitioner, ESORT: emergency surgery or not; TTO: tendency to operate

Figure 3.1 (b): Flowchart of eligibility for a target trial of emergency surgery versus non-emergency surgery for gallstone disease, emulation using Hospital Episodes Statistics data


2,310,797 admissions assessed for eligibility
1,986,155 Admissions failed to meet the inclusion criteria
1,716,899 Acute gallstone disease was not the reason for admission
201,796 Admission was not an emergency
78 Admission belonged to under-18 patient
298 Admission was to ineligible trust for ESORT
67,084 Appendicitis diagnosis was not confirmed by a surgeon
324,642 admissions met the inclusion criteria
83,665 admissions met the exclusion criteria
45,349 Admission preceded by another admission for the condition within the previous year
3,076 Admission to ineligible hospital for ESORT or for calculating TTO
691 Patient was transferred within hospitals before surgical assessment
456 Patient had surgery at a date prior to the date of surgical assessment
317 Admission lacked information on admission discharge status
33,776 Admission started before 1 April 2010 or after 31 December 2019
240,977 eligible admissions

A&E: accident and emergency, GP: general practitioner, ESORT: emergency surgery or not; TTO: tendency to operate

Table 3.2. Patient characteristics of the two cohorts of patients by emergency surgery (ES) and non-emergency surgery (NES) groups

	Acute appendicitis (N=268,144)		**Acute gallstone disease (N=240,977)**	
	ES (n=247,506)	NES (n=20,638)	ES (n=52,004)	NES (n=188,973)
Gender: n (%)				
Male	134,270 (54)	10,409 (50)	15,140 (29)	63,046 (33)
Female	113,224 (46)	10,228 (50)	36,864 (71)	125,927 (67)
Age: mean	38 (16)	47 (20)	51 (18)	56 (19)
IMD quintile: n (%)				
1 – Most deprived	49,495 (20)	4,319 (21)	11,774 (23)	44,650 (24)
2	47,818 (20)	3,898 (19)	9,586 (19)	34,792 (19)
3	49,203 (20)	4,128 (20)	10,641 (21)	37,561 (20)
4	50,337 (21)	4,024 (20)	10,881 (21)	39,759 (21)
5 – Least deprived	46,636 (19)	3,907 (20)	8,686 (17)	30,285 (16)
SCARF index: n (%)				
Fit	206,796 (84)	15,015 (73)	34,056 (66)	114,973 (61)
Mild frailty	34,544 (14)	4,052 (20)	13,608 (26)	52,629 (28)
Moderate frailty	5,041 (2)	1,155 (6)	3,385 (6)	16,175 (9)
Severe frailty	1,125 (0)	416 (2)	955 (2)	5,196 (3)
Ethnicity: n (%)				
Black/Black mixed	5,771 (2)	627 (3)	827 (2)	3,923 (2)
Asian/Asian mixed	11,592 (5)	1,122 (5)	2,204 (4)	9,124 (5)
White	194,968 (79)	16,371 (79)	44,396 (85)	162,727 (86)
Chinese and other	9,054 (4)	708 (3)	997 (2)	4,092 (2)
Charlson index: n (%)				
0 – comorbidities	207,525 (84)	15,321 (74)	36,737 (71)	120,748 (64)
1	35,721 (14)	3,989 (19)	12,287 (24)	49,863 (26)
2	3,715 (2)	1,035 (5)	2,544 (5)	14,503 (8)
3+ – comorbidities	545 (0)	293 (1)	436 (1)	3,859 (2)

*SCARF: Secondary Care Administrative Records Frailty.

3.2.3.2 Cost-effectiveness results

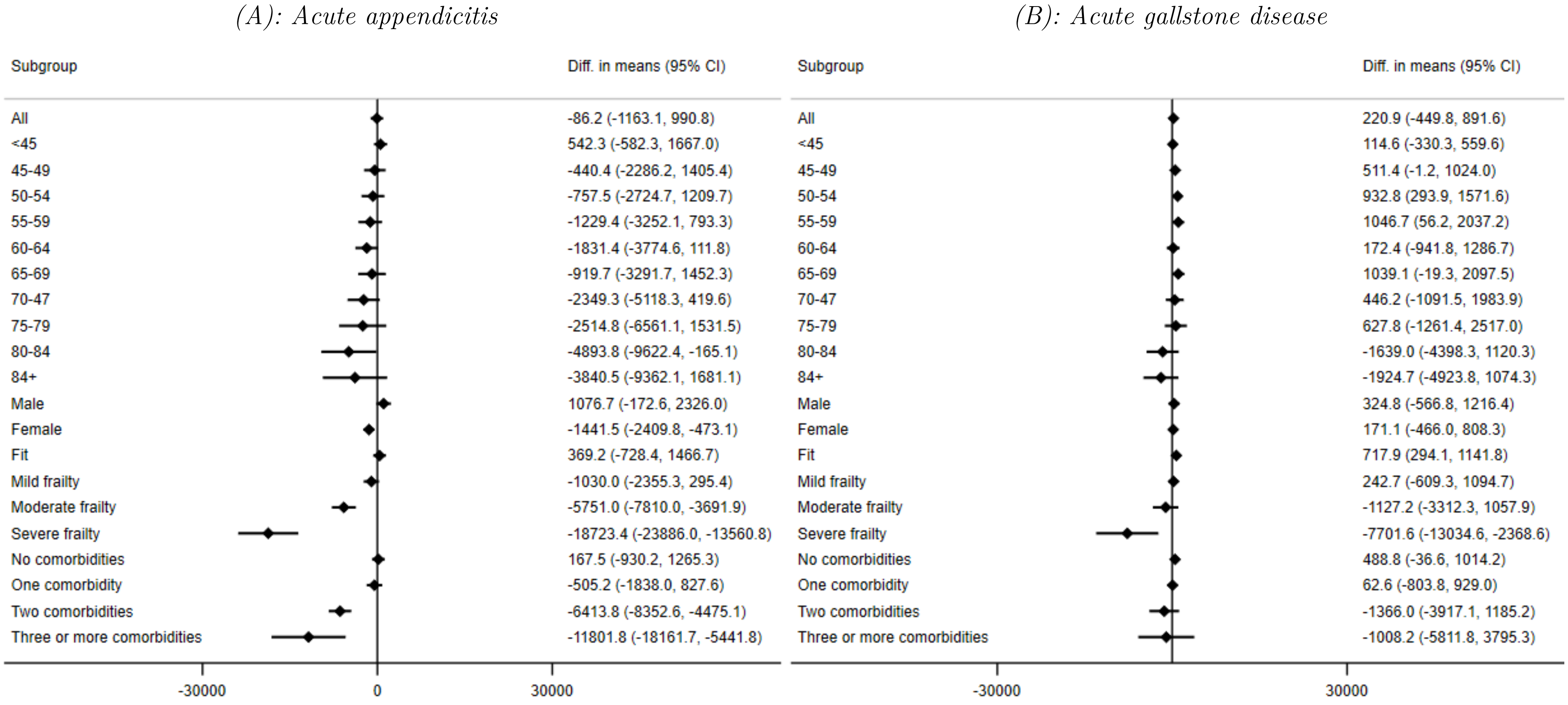
The LIV approach reports overall INB estimates for ES versus NES of -£86.2 (95% CI -1,163, 991) and £221 (-450, 892) for appendicitis and gallstone disease, respectively (Table 3.3). The regression adjustment reported similar estimates for the INB of -£223 (95% CI -342, -104) for acute appendicitis and -£220 (95% CI -316, 124) for gallstone disease (see also Appendix B.3 for estimated effects on costs, life years and QALYs). By contrast, the unadjusted INB estimates were £1,431 (95% CI 1,259, 1,603) and £1,002 (95% CI 832, 1,171) for acute appendicitis and gallstone disease, respectively (see also Appendix B.3 for estimated effects on costs, life years and QALYs). When considering population subgroups, the LIV analysis suggests that ES was not cost-effective for patients with severe frailty (for both conditions) and patients with two, three or more comorbidities (acute appendicitis) (Figure 3.2, see also Appendix B.4).

Table 3.3. Estimated group means and incremental costs (£GBP 2019/20), quality-adjusted life years (QALYs) and net monetary benefit (£GBP 2019/20, INB) at one year of emergency surgery vs non-emergency surgery strategies using the Local Instrumental Variable (LIV) approach

	Emergency surgery	**Non-emergency surgery**	**Mean differences (95% CI)**
	Acute appendicitis (N=268,144)		
Costs	3,366	3,475	-109 (-1,130, 913)
Life years	0.996	0.999	-0.003 (-0.006, -0.001)
QALYs	0.942	0.952	-0.010 (-0.024, 0.003)
Net benefit	15,475	15,561	-86.2 (-1,163, 991)
	Acute gallstone disease (N=240,977)		
Costs	5,477	5,554	-76.8 (-702, 548)
Life years	0.970	0.978	-0.009 (-0.022, 0.005)
QALYs	0.877	0.870	0.007 (-0.001, 0.015)
Net benefit	12,059	11,838	221 (-450, 892)

Variables used for adjustment in models: age (years), sex, ethnicity, index of multiple deprivation (quintiles), number of comorbidities (Charlson index), frailty level (SCARF index), method of admission, year fixed effects, proxies for the quality of acute care within the hospital.

Figure 3.2. Forest plots of estimated incremental net monetary benefit (INB) of emergency surgery (ES) versus non-emergency surgery (NES) for acute appendicitis (panel A) and acute gallstone disease (panel B) across population subgroups



(A): Acute appendicitis

Subgroup	Diff. in means (95% CI)
All	-86.2 (-1163.1, 990.8)
<45	542.3 (-582.3, 1667.0)
45-49	-440.4 (-2286.2, 1405.4)
50-54	-757.5 (-2724.7, 1209.7)
55-59	-1229.4 (-3252.1, 793.3)
60-64	-1831.4 (-3774.6, 111.8)
65-69	-919.7 (-3291.7, 1452.3)
70-47	-2349.3 (-5118.3, 419.6)
75-79	-2514.8 (-6561.1, 1531.5)
80-84	-4893.8 (-9622.4, -165.1)
84+	-3840.5 (-9362.1, 1681.1)
Male	1076.7 (-172.6, 2326.0)
Female	-1441.5 (-2409.8, -473.1)
Fit	369.2 (-728.4, 1466.7)
Mild frailty	-1030.0 (-2355.3, 295.4)
Moderate frailty	-5751.0 (-7810.0, -3691.9)
Severe frailty	-18723.4 (-23886.0, -13560.8)
No comorbidities	167.5 (-930.2, 1265.3)
One comorbidity	-505.2 (-1838.0, 827.6)
Two comorbidities	-6413.8 (-8352.6, -4475.1)
Three or more comorbidities	-11801.8 (-18161.7, -5441.8)

(B): Acute gallstone disease

Subgroup	Diff. in means (95% CI)
All	220.9 (-449.8, 891.6)
<45	114.6 (-330.3, 559.6)
45-49	511.4 (-1.2, 1024.0)
50-54	932.8 (293.9, 1571.6)
55-59	1046.7 (56.2, 2037.2)
60-64	172.4 (-941.8, 1286.7)
65-69	1039.1 (-19.3, 2097.5)
70-47	446.2 (-1091.5, 1983.9)
75-79	627.8 (-1261.4, 2517.0)
80-84	-1639.0 (-4398.3, 1120.3)
84+	-1924.7 (-4923.8, 1074.3)
Male	324.8 (-566.8, 1216.4)
Female	171.1 (-466.0, 808.3)
Fit	717.9 (294.1, 1141.8)
Mild frailty	242.7 (-609.3, 1094.7)
Moderate frailty	-1127.2 (-3312.3, 1057.9)
Severe frailty	-7701.6 (-13034.6, -2368.6)
No comorbidities	488.8 (-36.6, 1014.2)
One comorbidity	62.6 (-803.8, 929.0)
Two comorbidities	-1366.0 (-3917.1, 1185.2)
Three or more comorbidities	-1008.2 (-5811.8, 3795.3)

* Values to the left (right) of the 0 axis denote that NES (ES) is cost-effective for the subgroup.

3.2.4 Discussion

International HTA agencies are expanding their use of comparative effectiveness evidence from RWD studies (Garrison et al., 2007; Makady et al., 2017). NICE's new real-world evidence framework sets out recommendations to help RWD studies provide trustworthy evidence to inform decision-making, which include using the target trial framework to inform study design choices (NICE, 2022). This paper illustrates how this framework can be applied to HTA in a study evaluating the cost-effectiveness of ES for two common acute gastrointestinal conditions, which exemplifies common challenges in applying the target trial alongside RWD to inform HTA. In Table 3.4, we draw on the findings from this study to outline some recommendations for future studies looking to assess comparative effectiveness from RWD.

This paper makes three important contributions to the literature. First, it contributes to the literature of methods for informing HTA decision-making with robust effectiveness evidence from RWD. NICE describes three main barriers to the adoption of real-world evidence in their evaluations: (i) the risk of bias, (ii) the quality and relevance of the data, and (iii) concerns about the trustworthiness of the evidence (NICE, 2022). To tackle concerns about the trustworthiness of evidence, study design choices need to be made traceable and transparent for decision-makers. Current good-practice recommendations, including the reporting of checklists for economic evaluations, provide, in general, insufficient basis for judging study design choices outside of RCTs (Faria et al., 2015; Orsini et al., 2020). The target trial framework allows users of the evidence generated from RWD to assess its rigorousness and trustworthiness according to how closely the study design mimics that of an RCT. Published RCTs estimates can be used as 'benchmarks' in HTA to assess choices about aspects of the study design, including the plausibility of the assumptions underlying the different statistical approaches (Franklin et al., 2021). A further step would be to use the target trial framework in the design of systematic reviews and network meta-analyses of RCTs (Zhao et al., 2020). However, in many settings, RCT evidence for benchmarking is unavailable or unsuitable as it fails to include the target populations, comparators or endpoints of decision-making relevance. This study shows that RWD can still be used to support HTA decision-making in those settings. While applying the notions of target trial framework helps ensure that groups are comparable, thereby reducing the potential for confounding, this study highlights the importance of considering statistical methods that make alternative underlying assumptions about residual confounding. In ESORT, the unadjusted comparison of means which makes the implausible assumption of no confounding at all leads to a different conclusion to the GLM regression and LIV approaches which make more plausible assumptions about confounding, and lead to similar results.

Second, the paper tackles the lack of guidance on how to apply the principles of the target trial framework in RWD studies to ensure they meet the main requirements of HTA. We identify a series of challenges that are raised when using routine data for emulating target trials pertaining to: (i) defining the study population, (ii) defining the intervention and all relevant comparator strategies, (iii) establishing time zero, and (iv) using appropriate methods to adjust for confounding. Table 4 offers point-by-point recommendations for how to address these challenges.

The first challenge relates to the inability to emulate the target trial's eligibility criteria, which can result in bias due to imbalances in the distribution of patient characteristics. To inform HTA, applying the target trial framework would require RWD studies to emulate trials with active comparators (the 'standard of care') (NICE, 2014). Then, in order to minimise the risk of confounding from imbalances in prognostic factors, the eligibility criteria need to ensure that only patients for whom there is likely to be equipoise between treatment strategies are included. In the ESORT study, the criterion that the patient must be 'under the care of a surgeon' (see Table 3.1 for definition) helped exclude patients whose prognosis was so poor according to unobserved, as well as observed characteristics, that they would not be considered for ES (e.g., patients in advanced stages of the disease). When defining the eligibility criteria, another important consideration is that the population needs to include all patient subgroups of relevance for HTA decision-making. When published clinical guidance is insufficient to identify these populations, expert judgement should be used to adapt the target trial's eligibility criteria to the data available and to the requirements of HTA (see Table 3.4). Sensitivity analyses around the different eligibility criteria could help assess the implications of these decisions and should be adopted (Lodi et al., 2019). In the ESORT study, the clinical panel exercise provided a basis for this. The study could define alternative more/less strict definitions of the eligibility criteria by varying the threshold for required number of responses favouring inclusion of an ICD-10 code sub-category.

Table 3.4. Challenges and recommendations for studies applying the target trial framework alongside Real-World Data (RWD) to inform Health Technology Assessment (HTA)

Protocol	Challenge for RWD	Implications for HTA decision-making	Example from target trial of ES	Recommendation
Eligibility criteria	Data might be insufficient to emulate the trial's eligibility criteria	Estimates of comparative effectiveness could be subject to selection bias/ confounding if the distributions of patient characteristics are not balanced	Unclear which ICD-10 diagnostic subcategories describe patients with diagnoses of acute appendicitis and acute gallstone disease.	Use expert opinion to adapt the trial's eligibility criteria to the data available
	Population selected for study might include patients for whom there is no equipoise between treatment strategies	Estimates of comparative effectiveness could be subject to confounding bias	No equipoise for some patients with designated diagnostic codes for the condition (e.g., pregnant patients with designated codes of appendicitis)	Use clinical guidelines and/or expert opinion to define and exclude patient subgroups for whom there is no equipoise
	Population selected for study might fail to include subgroups of interest for decision-making	Findings could fail to inform decision-making if they are not generalisable to the target population, or omit relevant subgroup analyses	Unclear which patients are eligible and in equipoise for ES and NES strategies in routine practice	Use clinical guidelines and/or expert opinion to define subgroups of interest
Treatment strategies	The definition of the intervention (e.g., its timing) might differ from the intervention of interest	Findings could fail to inform decision-making if they do not reflect routine clinical practice	Unclear which OPCS-4 procedure codes and timings describe ES.	Use clinical guidelines and/or expert opinion to define the intervention and comparators
	The comparator strategy might not be defined with sufficient level of detail	Findings could fail to inform decision-making due to the interventions involved in the causal contrast not being well defined	The study could not inform the comparative effectiveness of ES versus specific NES treatments, but could do so against not receiving ES.	Carefully assess whether the causal contrast can be estimated given the data available.

Protocol	Challenge for RWD	Implications for HTA decision-making	Example from target trial of ES	Recommendation
Time zero	Start of follow-up might pre-date the assessment of the eligibility criteria	Findings could be subject to selection bias	Using the date of admission as day zero could result in bias due to post-baseline events being used to exclude patients.	Consider the likely bias arising from alternative candidates for day zero.
	Time of treatment assignment might not be aligned with that of eligibility assessment and start of follow-up	Findings could be subject to immortal time bias	Using the date of admission as day zero could result in bias if, during time until treatment initiation, the risk of event of interest differed between the groups.	Include as a criterion for day zero that it should minimise time to treatment initiation.
Statistical analysis	Residual confounding might exist after emulating the main components of the target trial, from both measured and unmeasured prognostic factors.	Estimates of comparative effectiveness could be biased by residual confounding	Naïve comparisons are unlikely to provide robust estimates, whereas adjustment in LIV and GLM regression resulted in similar findings	Consider appropriate methods for tackling confounding and, where possible, assess the underlying assumptions in the method used.
	Not all statistical methods might be appropriate for studying the causal contrast(s) of interest.	Findings might not be generalisable to the target population	Estimates of traditional IV methods usually pertain to narrow populations, but LIV can retrieve an overall effect.	Carefully assess the plausibility of the assumptions required for the estimation of the causal contrast.

The main challenge in defining the intervention and comparator strategies is to specify the treatment(s), dosage(s) and/or timing(s) that characterise their provision in routine clinical practice (second challenge). The definition could be informed by clinical guidelines for management of the condition, but as in the ESORT study, these are often unavailable. Unless the treatments of interest are specified within the RWD, the study will be of limited use for informing HTA decision-making (Hernán, 2004). Further to this, the study should carefully consider whether the comparators are defined in sufficient detail to evaluate the causal contrast of interest (Hernán and Robins, 2020; Holland, 1986). We recommend drawing on expert opinion to define the interventions and comparators of interest from those recorded within the routine data (Table 3.4).

The ESORT study highlights the challenges in defining time zero (baseline) from the RWD (third challenge), which cannot precede eligibility, and must minimise any delay prior to treatment initiation. In studies like ESORT, where treatment initiation for one or all treatment strategies is not observed in the data, the choice of time zero should be carefully evaluated. The ESORT study defined time zero as the date when the patient was first under the care of a surgeon. This definition is expected to carry low risk of bias since, (i) it is does not precede the time of eligibility assessment and, (ii) while it may not coincide with the time of treatment initiation, the probability of events until treatment initiation is small for these conditions. To help ensure the definition of time zero meets the requirement above, tools that help establish temporality from RWD, such as design diagrams (Patorno et al., 2020), and approaches like reweighting, censoring and cloning (Hernán and Robins, 2016) should be adopted in settings where immortal time bias is suspected.

In relation to the fourth challenge, our paper builds on precedent work on the use of IV methods for confounding adjustment, and in particular the combination of the target trial framework with IV methods to reduce the risk of bias from unmeasured confounding, which is a major concern in RWD (Swanson, 2017). ESORT uses a LIV approach which, unlike traditional IV methods such as two-stage least squares, can provide estimates of the ATE and CATEs that apply directly to the target population.

The application of the target trial framework should encompass the use of design tools pertaining to the choice of statistical approaches for estimating treatment effects in observational studies, such as the STROBE checklist (Vandenbroucke et al., 2007), and as this paper illustrates, a fundamental element of this is that the plausibility of the underlying assumptions is assessed, and alternative approaches that make contrasting assumptions are considered.

The third contribution of this paper is to illustrate how the target trial framework can be applied and used to make treatment recommendations in settings where appropriate RCT evidence is not available. The assessment of the relative cost-effectiveness of ES for acute appendicitis and acute gallstone disease in ESORT contributes to the scarce evidence on the effects of providing ES versus alternative strategies for patients with acute gastrointestinal conditions. For these conditions, studies conducted so far have evaluated ES against NES in relation to patient outcomes like mortality, HRQoL and length of hospital stay (Flum et al., 2020; Hutchings et al., 2022), but this paper contributes to the limited available evidence on relative cost-effectiveness (Javanmard-Emamghissi et al., 2020). In particular, while the ESORT study finds that overall, it is highly uncertain whether ES is cost-effective for treating patients with these three conditions, the results clearly suggest that for patients who have severe frailty ES is not cost-effective. This finding has direct implications for clinical decision-making, emphasises the importance of perioperative frailty assessment for patients presenting with these common conditions, and that alternative NES strategies including medical management or later surgery are more cost-effective for these patients

While the ESORT study exemplifies key issues that arise in undertaking emulations of target trials for HTA using individual patient data from routine sources, it cannot consider all the issues that may arise when using RWD in HTA. In ESORT, given the completeness and accuracy of HES data (ESORT Study Group, 2020c), there were no concerns around the risk of attrition bias or reporting bias, which can result from imbalances in the duration of follow-up and reporting of outcome data, but could be present in other studies. A related limitation of this study is that the application of the target trial framework was to the endpoints available within the routine data, namely survival time and health service utilisation. In other settings, lack of data on broader outcome measures could add another layer of complexity to the study. Finally, the ESORT study directly addresses the use of RWD for HTA purposes when individual patient data are available from a single study. More generally, greater consideration is needed on how the principles may expand to settings where individual patient data are not available for any, or all the comparators of interest (e.g., creating external controls in single arms trials).

In conclusion, this paper addresses common challenges that arise when applying the target trial framework to assess comparative effectiveness and cost-effectiveness for the purposes of HTA, when using RWD. The paper provides recommendations for improving the study design pertaining to the definition of the study population, comparators, and analytical approaches to help address concerns about the use of RWD in decision-making.

References

Bell H, Wailoo AJ, Hernandez M, et al. (2016) *The use of real world data for the estimation of treatment effects in NICE decision making: Report by the Decision Support Unit.*

Caniglia EC, Zash R, Jacobson DL, et al. (2018) Emulating a target trial of antiretroviral therapy regimens started before conception and risk of adverse birth outcomes. *AIDS* 32(1): 113–120. DOI: 10.1097/QAD.0000000000001673.Emulating.

Di Saverio S, Podda M, De Simone B, et al. (2020) Diagnosis and treatment of acute appendicitis: 2020 update of the WSES Jerusalem guidelines. *World Journal of Emergency Surgery* 15(1). World Journal of Emergency Surgery: 1–42. DOI: 10.1186/s13017-020-00306-3.

Dickerman BA, García-Albéniz X, Logan RW, et al. (2019) Avoidable flaws in observational analyses: an application to statins and cancer. *Nature Medicine* 25(10). Springer US: 1601–1606. DOI: 10.1038/s41591-019-0597-x.

Emilsson L, García-Albéniz X, Logan RW, et al. (2018) Examining bias in studies of statin treatment and survival in patients with cancer. *JAMA oncology* 4(1). American Medical Association: 63–70.

ESORT Study Group (2020a) Emergency Surgery Or NoT (ESORT) study. Available at: https://www.lshtm.ac.uk/media/38711.

ESORT Study Group (2020b) Emergency Surgery Or NoT (ESORT) study. Available at: https://www.lshtm.ac.uk/media/39151.

ESORT Study Group (2020c) Emergency Surgery Or NoT (ESORT) study. Available at: https://www.lshtm.ac.uk/media/51011.

Faria R, Hernández Alava M, Manca A, et al. (2015) *NICE DSU Technical Support Document 17: The use of observational data to inform estimates of treatment effectiveness in technology appraisal: Methods for comparative individual patient data.*

Flum DR, Davidson GH, Monsell SE, et al. (2020) A Randomized Trial Comparing Antibiotics with Appendectomy for Appendicitis. *New England Journal of Medicine* 383(20): 1907–1919. DOI: 10.1056/nejmoa2014320.

Franklin JM, Patorno E, Desai RJ, et al. (2021) Emulating Randomized Clinical Trials with Nonrandomized Real-World Evidence Studies: First Results from the RCT DUPLICATE Initiative. *Circulation* (143): 1002–1013. DOI: 10.1161/CIRCULATIONAHA.120.051718.

Freemantle N, Marston L, Walters K, et al. (2013) Making inferences on treatment effects from real world data: propensity scores, confounding by indication, and other perils for the unwary in observational research. Available at: http://dx.doi.org/10.1136/bmj.f6409.

Garrison LP, Neumann PJ, Erickson P, et al. (2007) Using real-world data for coverage and payment decisions: The ISPOR real-world data Task Force report. *Value in Health* 10(5). International Society for Pharmacoeconomics and Outcomes Research (ISPOR): 326–335. DOI: 10.1111/j.1524-4733.2007.00186.x.

Gomes M, Latimer N, Soares M, et al. (2022) Target trial emulation for transparent and robust estimation of treatment effects for health technology assessment using real-world data: opportunities and challenges. *PharmacoEconomics*. Springer: 1–10.

Hernán MA (2004) A definition of causal effect for epidemiological research. *Journal of Epidemiology and Community Health* 58(4): 265–271. DOI: 10.1136/jech.2002.006361.

Hernán MA and Robins JM (2016) Using Big Data to Emulate a Target Trial When a Randomized Trial Is Not Available. *American Journal of Epidemiology* 183(8). Oxford University Press: 758–764. DOI: 10.1093/aje/kwv254.

Hernán MA and Robins JM (2020) *Causal Inference: What If*. Boca Raton: Chapman & Hall/CRC.

Hernán MA and Taubman SL (2008) Does obesity shorten life? The importance of well-defined interventions to answer causal questions. *International Journal of Obesity* 32: S8–S14. DOI: 10.1038/ijo.2008.82.

Hernán MA, Sauer BC, Hernández-Díaz S, et al. (2016) Specifying a target trial prevents immortal time bias and other self-inflicted injuries in observational analyses. *Journal of Clinical Epidemiology* 79: 70–75. DOI: 10.1016/j.jclinepi.2016.04.014.Specifying.

Holland PW (1986) Statistics and causal inference. *Journal of the American statistical Association* 81(396). Taylor & Francis: 945–960.

Husereau D, Drummond M, Petrou S, et al. (2013) Consolidated Health Economic Evaluation Reporting Standards (CHEERS) statement. *European Journal of Health Economics* 14: 367–372. DOI: 10.1007/s10198-013-0471-6.

Hutchings A, O'Neill S, Lugo-palacios DG, et al. (2022) Effectiveness of emergency surgery for five common acute conditions: an instrumental variable analysis of a national routine database. *Anaesthesia*: In Press.

Javanmard-Emamghissi H, Boyd-Carson H, Hollyman M, et al. (2020) The management of adult appendicitis during the COVID-19 pandemic: an interim analysis of a UK cohort study. *Techniques in Coloproctology* (0123456789). DOI: 10.1007/s10151-020-02297-4.

Lévesque LE, Hanley JA, Kezouh A, et al. (2010) Problem of immortal time bias in cohort studies: Example using statins for preventing progression of diabetes. *BMJ (Online)* 340(7752): 907–911. DOI: 10.1136/bmj.b5087.

Lodi S, Phillips A, Lundgren J, et al. (2019) Effect Estimates in Randomized Trials and Observational Studies: Comparing Apples with Apples. *American Journal of Epidemiology* 188(8). Oxford Academic: 1569–1577. DOI: 10.1093/aje/kwz100.

Makady A, Ham R ten, de Boer A, et al. (2017) Policies for Use of Real-World Data in Health Technology Assessment (HTA): A Comparative Study of Six HTA Agencies. *Value in Health* 20(4). Elsevier Inc.: 520–532. DOI: 10.1016/j.jval.2016.12.003.

Maringe C, Benitez Majano S, Exarchakou A, et al. (2020) Reflection on modern methods: Trial emulation in the presence of immortal-time bias. Assessing the benefit of major surgery for elderly lung cancer patients using observational data. *International Journal of Epidemiology* 49(5): 1719–1729. DOI: 10.1093/ije/dyaa057.

Moler-Zapata S, Grieve R, Lugo-Palacios D, et al. (2022) Local instrumental variable methods to address confounding and heterogeneity when using electronic health records: an application to emergency surgery. *Medical Decision Making* 0(0). DOI: 10.1177/0272989X221100799.

National Confidential Enquiry into patient outcomes and death (2004) *The NCEPOD Classification of Intervention.* Available at: www.ncepod.org.uk/classification.html (accessed 3 July 2022).

National Institute for Health and Care (2013) Guide to the methods of technology appraisal. London. Available at: https://www.nice.org.uk/process/pmg9/chapter/foreword.

National Institute for Health and Care Excellence (2014) Developing NICE guidelines: the manual. *Process and methods guides* (October): 245. Available at: http://www.nice.org.uk/article/pmg20.

National Institute for Health and Care Excellence (2022) *NICE real-world evidence framework*. London (UK). Available at: www.nice.org.uk/corporate/ecd9.

Orsini LS, Monz B, Mullins CD, et al. (2020) Improving transparency to build trust in real-world secondary data studies for hypothesis testing—Why, what, and how: recommendations and a road map from the real-world evidence transparency initiative. *Pharmacoepidemiology and Drug Safety* 29(11): 1504–1513. DOI: 10.1002/pds.5079.

Patorno E, Schneeweiss S and Wang S V. (2020) Transparency in real-world evidence (RWE) studies to build confidence for decision-making: Reporting RWE research in diabetes. *Diabetes, Obesity and Metabolism* 22(S3): 45–59. DOI: 10.1111/dom.13918.

Petersen, M. L., Porter, K. E., Gruber, S., Wang, Y., & van der Laan, M. J. (2012). Diagnosing and responding to violations in the positivity assumption. Statistical methods in medical research, 21(1), 31–54. https://doi.org/10.1177/0962280210386207

Petito LC, García-Albéniz X, Logan RW, et al. (2020) Estimates of Overall Survival in Patients With Cancer Receiving Different Treatment Regimens: Emulating Hypothetical Target Trials in the Surveillance, Epidemiology, and End Results (SEER)-Medicare Linked Database. *JAMA network open* 3(3): e200452. DOI: 10.1001/jamanetworkopen.2020.0452.

Swanson SA (2017) Instrumental Variable Analyses in Pharmacoepidemiology: What Target Trials Do We Emulate? *Current Epidemiology Reports* 4(4). Current Epidemiology Reports: 281–287. DOI: 10.1007/s40471-017-0120-1.

Thorn JC, Ridyard CH, Hughes D, et al. (2016) Health economics analysis plans: Where are we now? *Value in Health* 19(7). Elsevier.

Vandenbroucke JP, Elm E von, Altman DG, et al. (2007) Strengthening the Reporting of Observational Studies in Epidemiology (STROBE): explanation and elaboration. *Annals of internal medicine* 147(8). American College of Physicians: W-163.

Zhao SS, Lyu H, Solomon DH, et al. (2020) Improving rheumatoid arthritis comparative effectiveness research through causal inference principles: systematic review using a target trial emulation framework. *Annals of the Rheumatic Diseases 79(7).* BMJ Publishing Group: 1–24. DOI: 10.1136/ANNRHEUMDIS-2020-217200.

Chapter 4. Local instrumental variable methods to address confounding and heterogeneity when using electronic health records: an application to emergency surgery

4.1 Preamble to research paper

This chapter presents an application of the LIV methodology to a CEA using routine data from England. The paper was published in *Medical Decision Making* by open access on May 24, 2022, as part of the special theme issue on "The use of electronic health record (EHR) data in health decision research". The full reference for the article is:

Moler-Zapata S, Grieve R, Lugo-Palacios D, et al. (2022) Local instrumental variable methods to address confounding and heterogeneity when using electronic health records: an application to emergency surgery. *Medical Decision Making* 0(0). DOI: 10.1177/0272989X221100799.

Prior to this work, the LIV methodology developed by Heckman and Vytlacil (1999, 2001, 2005) and further extended by Basu (2014) had not been used in a CEA. This research, conducted within the ESORT study, sought to address the gap in the evidence on the relative the benefits, risk and costs of ES compared to alternative NES strategies for treating patients with common acute gastrointestinal conditions who are admitted into hospital as an emergency. This setting exemplifies how LIV methods can be used to expand the evidence base with real-world evidence (e.g., by considering broader study populations). Published RCTs for some of these conditions have, included highly selective patient samples, reported outcomes over short follow-up periods or failed to consider economic outcomes of relevance for policy-makers and health care providers such as resource use and costs (Azhar et al., 2021; Flum et al., 2020; Javanmard-Emamghissi et al., 2021). For other acute conditions, such as abdominal wall hernia, no RCTs of ES have been conducted. Some published studies have had non-experimental designs but they have failed to address the fundamental concern of unmeasured confounding (Koumarelas et al., 2014; Saverio et al., 2014).

The study describes the target estimand and main assumptions required for identification with LIV. The paper illustrates how LIV can be used to evaluate heterogeneity of treatment effects over population groups, it also contrasts LIV against alternative IV approaches which make alternative assumptions, and offers guidance for future CEA on how to interpret any discrepancies between the different methods. My role included designing the CEA, collating resource use data from HES, collating unit cost data from national databases, conducting literature searches to identify HRQoL data, evaluating the identification assumptions, and conducting the LIV analyses, jointly with my supervisor, SON. I led the interpretation of the results. I wrote the draft version of the manuscript, and incorporated comments from co-authors into the manuscript. I also addressed the comments raised during the peer-review process.

The analysis received ethical approval from the LSHTM Ethics Committee (ID:21776).

References

Azhar N, Johanssen A, Sundström T, et al. (2021) Laparoscopic Lavage vs Primary Resection for Acute Perforated Diverticulitis: Long-term Outcomes From the Scandinavian Diverticulitis (SCANDIV) Randomized Clinical Trial. *JAMA Surgery* 156(2): 121–128. DOI: 10.1001/jamasurg.2020.5618.

Basu A (2014) Estimating person-centered treatment (PeT) effects using instrumental variables: an application to evaluating prostate cancer treatments. *Journal fo applied econometrics* 29: 671–691. DOI: 10.1002/jae.

Flum DR, Davidson GH, Monsell SE, et al. (2020) A Randomized Trial Comparing Antibiotics with Appendectomy for Appendicitis. *New England Journal of Medicine* 383(20): 1907–1919. DOI: 10.1056/nejmoa2014320.

Heckman JJ and Vytlacil E (2005) Structural equations, treatment effects and econometric policy evaluation. *Econometrica* 73(3): 669–738. DOI: 10.1111/j.1468-0262.2005.00594.x.

Heckman JJ and Vytlacil EJ (1999) Local instrumental variables and latent variable models for identifying and bounding treatment effects. *Proceedings of the National Academy of Sciences of the United States of America* 96: 4730–4734. DOI: 10.1073/pnas.96.8.4730.

Heckman JJ and Vytlacil EJ (2001) Policy-Relevant Treatment Effects. *American Economic Review* 91(2): 107–111. DOI: 10.1257/aer.91.2.107.

Javanmard-Emamghissi H, Hollyman M, Boyd-Carson H, et al. (2021) Antibiotics as first-line alternative to appendicectomy in adult appendicitis: 90-day follow-up from a prospective, multicentre cohort study. *British Journal of Surgery*: 1–9. DOI: 10.1093/bjs/znab287.

Koumarelas K, Theodoropoulos GE, Spyropoulos BG, et al. (2014) A prospective longitudinal evaluation and affecting factors of health related quality of life after appendectomy. *International Journal of Surgery* 12(8). Elsevier Ltd: 848–857. DOI: 10.1016/j.ijsu.2014.06.015.

Saverio S Di, Sibilio AM, Giorgini E, et al. (2014) The NOTA Study (Non Operative Treatment for Acute Appendicitis): Prospective Study on the Efficacy and Safety of Antibiotics (Amoxicillin and Clavulanic Acid) for Treating Patients With Right Lower Quadrant Abdominal Pain and Long-Term Follow-up of Conser. *Annals of Surgery* 260(1): 109–117. DOI: 10.1136/bmjopen-2010-000006.

London School of Hygiene & Tropical Medicine
Keppel Street, London WC1E 7HT

T: +44 (0)20 7299 4646
F: +44 (0)20 7299 4656
www.lshtm.ac.uk

RESEARCH PAPER COVER SHEET

Please note that a cover sheet must be completed for each research paper included within a thesis.

SECTION A – Student Details

Student ID Number	1903225	**Title**	Ms
First Name(s)	Silvia		
Surname/Family Name	Moler Zapata		
Thesis Title	METHODS TO ADDRESS CONFOUNDING AND HETEROGENEITY IN COST-EFFECTIVENESS ANALYSES USING REAL-WORLD DATA		
Primary Supervisor	Prof Richard Grieve		

If the Research Paper has previously been published please complete Section B, if not please move to Section C.

SECTION B – Paper already published

Where was the work published?	Medical Decision Making		
When was the work published?	Augut 2022		
If the work was published prior to registration for your research degree, give a brief rationale for its inclusion	N/A		
Have you retained the copyright for the work?*	**Yes**	Was the work subject to academic peer review?	**Yes**

*If yes, please attach evidence of retention. If no, or if the work is being included in its published format, please attach evidence of permission from the copyright holder (publisher or other author) to include this work.

SECTION C – Prepared for publication, but not yet published

Where is the work intended to be published?	
Please list the paper's authors in the intended authorship order:	

Stage of publication	**Not yet submitted**

SECTION D – Multi-authored work

For multi-authored work, give full details of your role in the research included in the paper and in the preparation of the paper. (Attach a further sheet if necessary)	My role included designing the CEA, collating resource use data from HES, collating unit cost data from national databases, conducting literature searches to identify HRQoL data, evaluating the identification assumptions, and conducting the LIV analyses, jointly my supervisor, SON. I led the interpretation of the results. I wrote the draft version of the manuscript, and incorporated comments from co-authors into the manuscript. Upon acceptance of the paper for peer review, I addressed the comments raised during the revision process.

SECTION E

Student Signature	Silvia Moler Zapata
Date	23 Sep. 22

Supervisor Signature	Richard Grieve
Date	23 Sep. 22

4.2 Research paper 2: Local instrumental variable methods to address confounding and heterogeneity when using electronic health records: an application to emergency surgery

Authors

Silvia Moler-Zapata[1], Richard Grieve[1], David Lugo-Palacios[1], A. Hutchings[1], R. Silverwood[2], Luke Keele[3], Tommaso Kircheis[1], David Cromwell[1,4], Neil Smart[5], Robert Hinchliffe[6], and Stephen O'Neill[1].

Affiliations

[1]Department of Health Services Research and Policy, London School of Hygiene & Tropical Medicine, London, UK.
[2]University College London, London, UK.
[3]University of Pennsylvania, Philadelphia, USA. Associate Professor of Statistics.
[4]Clinical Effectiveness Unit, Royal College of Surgeons of England, London, UK.
[5]College of Medicine and Health, University of Exeter, Exeter, UK.
[6]Bristol Surgical Trials Centre, University of Bristol, Bristol, UK.

Abstract

Background: Electronic health records (EHRs) offer opportunities for comparative effectiveness research to inform decision making. However, to provide useful evidence, these studies must address confounding and treatment effect heterogeneity according to unmeasured prognostic factors. Local instrumental variable (LIV) methods can help studies address these challenges, but have yet to be applied to EHR data. This article critically examines a LIV approach to evaluate the cost-effectiveness of emergency surgery (ES) for common acute conditions from EHRs.

Methods: This article uses hospital episodes statistics (HES) data for emergency hospital admissions with acute appendicitis, diverticular disease, and abdominal wall hernia to 175 acute hospitals in England from 2010 to 2019. For each emergency admission, the instrumental variable for ES receipt was each hospital's ES rate in the year preceding the emergency admission. The LIV approach provided individual-level estimates of the incremental quality-adjusted life-years, costs and net monetary benefit of ES, which were aggregated to the overall population and subpopulations of interest, and contrasted with those from traditional IV and risk-adjustment approaches.

Results: The study included 268,144 (appendicitis), 138,869 (diverticular disease), and 106,432 (hernia) patients. The instrument was found to be strong and to minimize covariate imbalance. For diverticular disease, the results differed by method; although the traditional approaches reported that, overall, ES was not cost-effective, the LIV approach reported that ES was cost-effective but with wide statistical uncertainty. For all 3 conditions, the LIV approach found heterogeneity in the cost-effectiveness estimates across population subgroups: in particular, ES was not cost-effective for patients with severe levels of frailty.

Conclusions: EHRs can be combined with LIV methods to provide evidence on the cost-effectiveness of routinely provided interventions, while fully recognizing heterogeneity.

Keywords

Cost-effectiveness analysis, emergency surgery, heterogeneous treatment effects, instrumental variable, personalized medicine

Highlights

- This article addresses the confounding and heterogeneity that arise when assessing the comparative effectiveness from electronic health records (EHR) data, by applying a local instrumental variable (LIV) approach to evaluate the cost-effectiveness of emergency surgery (ES) versus alternative strategies, for patients with common acute conditions (appendicitis, diverticular disease, and abdominal wall hernia).
- The instrumental variable, the hospital's tendency to operate, was found to be strongly associated with ES receipt and to minimize imbalances in baseline characteristics between the comparison groups.
- The LIV approach found that, for each condition, there was heterogeneity in the estimates of cost-effectiveness according to baseline characteristics.
- The study illustrates how an LIV approach can be applied to EHR data to provide cost-effectiveness estimates that recognize heterogeneity and can be used to inform decision making as well as to generate hypotheses for further research.

4.2.1 Introduction

Electronic health records (EHRs) offer important opportunities for comparative effectiveness research that can directly inform medical decision making (Kuo et al., 2018; Russell, 2021). EHRs offer the possibility of evaluating interventions as provided in practice to all eligible patients. Agencies, such as the National Institute for Health and Care Excellence (NICE), recognize the potential of EHRs (NICE, 2013), but to provide useful evidence about comparative effectiveness, two major concerns must be addressed. First, treatment selection according to unmeasured baseline prognostic measures (e.g., disease severity) can make results subject to unmeasured confounding (Kreif et al., 2013; Kyriacou and Lewis, 2016). Second, there may be treatment effect heterogeneity according to patient and contextual characteristics. While approaches for handling heterogeneity according to measured covariates (effect modification) are commonly used, less attention has been given to 'essential heterogeneity', that is, heterogeneous gains according to unmeasured characteristics that influence selection into treatment (Basu et al., 2007; Heckman et al., 2006).

The first challenge is unlikely to be addressed by studies that apply traditional risk adjustment methods to provide estimates of comparative effectiveness, as EHRs tend to have inadequate information on case severity (Keele and Small, 2019; Stürmer et al., 2011). A valid instrumental variable (IV) design can provide accurate estimates of treatment effectiveness, even when there are unmeasured differences between the comparison groups (Baiocchi et al., 2014). If the IV is valid, it encourages receipt of the treatment, but does not have an effect on the outcome, except through treatment receipt. However, a major concern with applying traditional IV approaches, such as 2-stage least squares (2SLS) in the presence of essential heterogeneity, is that the resultant estimates are unlikely to apply to the overall populations or subpopulations of decision-making interest (Angrist et al., 1993; Angrist and Krueger, 1999; Baiocchi et al., 2014; Imbens and Angrist, 1994).

Local instrumental variable (LIV) approaches can provide estimates of comparative effectiveness that apply to policy-relevant populations (Heckman and Vytlacil, 1999, 2001, 2005). LIV methods can estimate individual-level treatment effects, known as person-centered treatment (PeT) effects, which can then be aggregated over relevant subgroups. LIV methods make the same underlying assumptions as all IV methods but also require that the instrument be continuous (Heckman and Vytlacil, 2005). LIV approaches have been used for comparative effectiveness research as part of bespoke observational studies of educational reforms (Basu, Jones, et al., 2018), cardiovascular and bariatric surgery (Coleman et al., 2020; Reynolds et al., 2021), and transfers to intensive care units (Grieve et al., 2019), but they have not been applied to EHR data, nor to an economic evaluation. In EHR settings, it is particularly challenging to identify and assess the validity of an IV, given that the data are collected for clinical or administrative rather than research purposes.

These major challenges of using EHRs for comparative effectiveness research are exemplified by the ESORT study (ESORT Study Group, 2020), which aims to evaluate the effectiveness and cost-effectiveness of ES versus nonemergency surgery (NES) strategies, which include antibiotic therapy, nonsurgical procedures (e.g., drainage of abscess), or surgery deferred to the elective (planned) setting. The question as to whether ES or NES strategies are more cost-effective is important, given the high burden of emergency general surgical services and the lack of evidence to inform clinical decision making (Abbott et al., 2017; Abercrombie, 2017; Stewart et al., 2014). Here, an unmet challenge is to identify those patient groups for whom ES is most cost-effective, and conversely those for whom NES alternatives, such as later surgery, may be more worthwhile. Randomized controlled trials (RCTs) have been undertaken for some acute conditions such as acute appendicitis and diverticular disease, but these have included highly selective or small patient samples, whereas for other acute conditions, such as abdominal wall hernia, no RCTs of ES have been conducted (Azhar et al., 2021; Flum et al., 2020; Javanmard-Emamghissi et al., 2021; Thornell et al., 2016).

Faced with this evidence gap, the ESORT study uses records from England's Hospital Episode Statistics (HES) database on emergency admissions to acute National Health Service (NHS) hospitals from 2009 to 2019, for common acute conditions, including the 3 considered in this article, acute appendicitis, diverticular disease, and abdominal wall hernia (ESORT Study Group, 2020) HES for admitted patient care is a database containing administrative, patient, and clinical details of all admissions to hospitals in England's NHS (Herbert et al., 2017). Clinical data on diagnoses and procedures are routinely extracted from discharge summaries for inclusion in local patient information databases, and transferred to HES. The HES database is primarily used for administrative and payment purposes. HES lacks detailed clinical data held locally but has been used widely for research purposes. The ESORT study previously used HES data and found no evidence of differences in the overall clinical effectiveness of ES versus NES strategies (Hutchings et al., 2022). However, this earlier article did not consider alternative approaches for tackling the confounding that arises with HES data or provide the estimates of relative cost-effectiveness that are essential for decision making.

The aim of this article is to critically examine LIV methods for addressing unmeasured confounding and heterogeneity in evaluating the cost-effectiveness of ES for patients with these 3 conditions from EHR data. The article is structured as follows. First, we provide an overview of the ESORT study. Second, we define the main aspects of the LIV methodology, including application to the ESORT study. Third, we present the results. Fourth, we discuss the key findings, strengths, and limitations of the article and the implications for further research.

4.2.2 Methods

4.2.2.1 Essential features of the ESORT study

Data sources and study population. The ESORT study uses HES data to evaluate the relative effectiveness and cost-effectiveness of ES versus alternative strategies from the hospital perspective over a 1-y time horizon. The study protocol and statistical analysis plan were developed following the principles of the target trial emulation framework (ESORT Study Group, 2020a; Hernán and Robins, 2016). Briefly, the ESORT study includes patients aged 18 y or older, admitted as an emergency admission via an accident and emergency department, or primary care referral, who were admitted to 175 NHS hospitals in England from April 1, 2010, to December 31, 2019; had the relevant ICD-10 diagnostic codes; and met other inclusion criteria (see Appendix C.6).

Comparator strategies. Admissions were defined as receiving the ES strategy if, according to Office of Population Censuses and Surveys (OPCS) codes, they had a relevant operative procedure within time windows designated by a clinical panel of 3 d (hernia), 7 d (appendicitis), or any time within the emergency admission (diverticular disease) (ESORT Study Group, 2020). The NES strategies included medical management, interventional radiology, and operative procedures that did not meet the ES criteria (see Appendix C.6).

Covariates. Baseline covariates were extracted from HES and included age, sex, ethnicity, the Index of Multiple Deprivation, the Charlson Comorbidity Index (Armitage and Van Der Meulen, 2010), the secondary care administrative records frailty (SCARF) index (Jauhari et al., 2020), and teaching hospital status. The SCARF index uses ICD-10 codes to define 32 deficits that cover functional impairment, geriatric syndromes, problems with nutrition, cognition and mood, and medical comorbidities, with severe frailty defined as the presence of 6 or more deficits. Information was taken from HES data to derive proxy measures of the quality of acute care in each hospital according to rates of 90-d all-cause mortality and emergency readmissions in preceding periods. Subgroups of interest were defined ex ante, drawing on clinical judgment to define those strata anticipated to modify the relative effectiveness and cost-effectiveness of ES. Subgroup definitions were based on the following baseline characteristics: age group, sex, Charlson comorbidity index, SCARF index, diagnostic subcategories, and year of admission.

Outcomes. The CEA took an intention-to-treat approach, whereby all patients contributed to the treatment group to which they were assigned at baseline, irrespective of the subsequent treatments received (e.g., planned or unplanned surgery). We reported the mean (95% confidence interval) incremental costs, quality-adjusted life-years (QALYs), and net monetary benefit (INB) at 1 y. Individual-level resource use was extracted from HES data for the index emergency admission and for all subsequent hospital readmissions up to the end of follow-up (death or December 31, 2019). Resource use included the length of the hospital stay, including time in intensive care units, and the use of diagnostic and operative procedures. Resource use items were combined with unit costs (£ GDP, 2019/20) to calculate total costs per patient (see section 1 and Appendix C.7., C.8., and C.9.). All unit costs were inflated to 2019–20 prices (£ GBP) using UK's GDP deflator published by HM Treasury (HM Treasury Department, 2020).

Survival time up to 1 y was calculated for all patients from HES records linked to the Office for National Statistics death data. Health-related quality of life (HRQoL) data were not available from HES, and so QALYs were calculated by combining the survival time with HRQoL estimates from the literature (see Appendix C.2, C.3, C.10 and C.11). We derived each patient's QALYs at 1 y using the area under the curve approach (Manca et al., 2005), which allowed HRQoL to decrease to baseline levels following an emergency readmission, but assumed that HRQoL levels recovered following hospital discharge. HRQoL levels were adjusted to reflect the patient's age and gender, and were assumed to be zero for patients who died over the follow-up period (see Appendix C.18) (Ara et al., 2017; Ara and Brazier, 2010). The study's cost-effectiveness metric was the INB of ES versus NES, calculated by multiplying the incremental QALYs by a NICE recommended willingness-to-pay threshold of £20,000 per QALY and subtracting from this the incremental cost (NICE, 2013).

We now present the main elements of the LIV design (in the following section). We then discuss how PeT effects, average treatment effect (ATE), and conditional ATEs (CATEs) were estimated using LIV and contrast the results against 2 alternative methods for estimating the ATE—2-stage residual inclusion and GLM regression—which make different assumptions about confounding and heterogeneity.

4.2.2.2 Instrumental Variable estimation

4.2.2.2.1 Overview

A valid instrument must be associated with treatment assignment (relevance condition) (i), the IV must be independent of unmeasured confounders (exchangeability condition) (ii), the IV must influence the outcomes only through treatment assignment (exclusion-restriction assumption) (iii), and the IV must have the same direction of effect on the probability of which treatment is received, irrespective of the level of the IV (monotonicity) (iv) (Angrist et al., 1993; Baiocchi et al., 2014). The most widely used IV approach, 2SLS, estimates the average treatment effect (ATE) when effects are homogeneous. If there are heterogeneous treatment effects, and the IV is binary, 2SLS reports a local ATE (LATE) or a weighted average of LATEs with a continuous IV (Angrist and Imbens, 1995; Cornelissen et al., 2016), requiring careful interpretation of the estimated effects in light of the LATE estimand.

Two-stage residual inclusion

2-stage residual inclusion (2SRI) is an IV approach that relies on concepts that support control function methods in an attempt to control for unmeasured confounding (Terza et al., 2008). This approach uses residuals from a first-stage regression for treatment assignment, in a second-stage outcome model (Terza et al., 2008). Unlike 2SLS, the 2SRI approach, when applied to a binary treatment, aims to estimate the ATE rather than LATEs. However, concerns have been raised that this approach may provide biased estimates of the ATE due to the necessity to extrapolate the residuals when constructing counterfactuals, and that it is sensitive to misspecification of the functional form underlying the residuals (Basu, Coe, et al., 2018). Here, we address the latter concern by using generalized residuals, which have been shown to minimize the bias in estimating the ATE (Basu, Coe, et al., 2018). Nonetheless, although 2SRI can, in some circumstances, provide accurate estimates of the ATE, it is not specifically recommended for exploring heterogeneity (Terza et al., 2008).

4.2.2.2.2 Estimating person-level effects using Local Instrumental Variable Methods

We also consider an LIV method that can estimate ATEs, subgroup effects, and personalized treatment effects, in the presence of unmeasured confounding and heterogeneity, and can extend to nonlinear outcomes such as costs and QALYs (Basu, 2014; Basu et al., 2007).

Heckman and Vytlacil (1999, 2001, 2005) showed that LIV methods can identify effects for "marginal" patients, those who are in equipoise with respect to the treatment assignment decision, provided a valid, continuous instrument is available. These individuals' propensity for treatment (PS), based on the levels of their observed covariates and IV, just balance with a normalized version of the unmeasured confounders (*V*) discouraging treatment, such that a small (marginal) change in the IV is sufficient to nudge them into the treatment group (where D=1 [i.e., ES] if PS > V and 0 [NES] otherwise). Contrasting outcomes for individuals with marginally different values of the IV, but who are otherwise identical in measured and unmeasured covariates at different levels of the IV, identifies a series of marginal treatment effects (MTEs). The MTE is equivalent to the conditional LATE for infinitesimally small changes in the normalized unobserved confounder, *V* (Huber and Wüthrich, 2019). MTEs can then be aggregated to obtain the ATE and CATEs for subgroups (Heckman and Vytlacil, 2005).

The LIV method relies on correctly modeling the relationships of the covariates and the IV with both the treatment and the outcome, typically using parametric models (Kennedy et al., 2019; Ogburn et al., 2015). If the treatment assignment model is misspecified, the second-stage model will use biased estimates of the PS, thus introducing bias into the subsequent effect estimates. Similarly, if the outcome model is misspecified, the estimated MTEs may not represent the true MTEs, as they will have been derived as the derivative of an incorrect outcome model $MTE = \frac{\partial E(Y|X=x,Z=z)}{\partial ps}$. While the "true" model specifications are unknown, considering alternative specifications, visually inspecting the models' predictions versus actual values, and considering the root mean squared error (rMSE) of the predictions, in addition to using standard model diagnostic approaches such as Hosmer and Lemeshow (2000) and Pregibon (1980) tests for Generalised Linear Models (GLMs), can be helpful in minimising risk of misspecification.

Basu (2014) extended the LIV approach by using the individual patient's observed treatment status to obtain personalized effect estimates. The key insight underlying this approach is that for each individual patient, some levels of the normalized unobserved confounder would be inconsistent with the observed treatment decision for that individual, given their observed characteristics and the level of the IV (Basu, 2014). For instance, if an individual with high propensity for ES according to observables (e.g., age) were observed to receive NES, it is reasonable to assume that the discouragement according to unobserved confounders must have exceeded the propensity for ES (i.e. PS < V if D=0). MTEs that imply a lower level of unobserved confounding can thus be ‘ruled out’, narrowing the set of MTEs which could plausibly represent the individual’s effect. The person-centered treatment (PeT) effect for an individual is obtained by aggregating the remaining MTEs and, are therefore more nuanced or ‘personalized’ than MTEs and CATEs. These effects can then be aggregated to obtain higher level estimands (e.g., ATE and CATEs (Basu, 2014, 2015)). (For full details and implementation in this study, see Appendix C.4.).

4.2.2.2.3 Developing IV and LIV approaches within the ESORT study

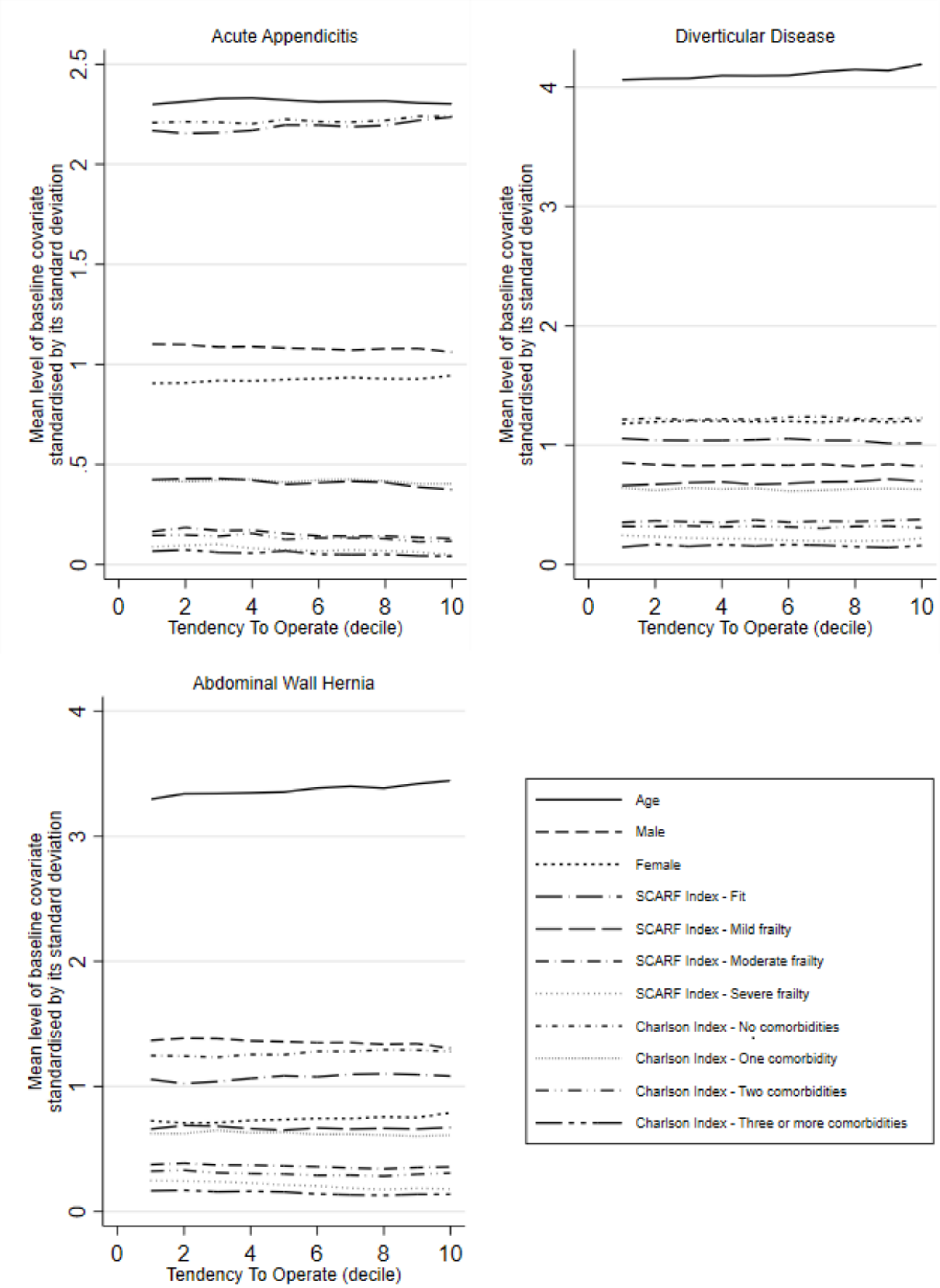
The ESORT study adopted an IV approach to evaluate ES from US claims data (Keele et al., 2018), following pharmaco-epidemiological research in taking clinician preference as an instrument for treatment receipt (Brookhart and Schneeweiss, 2007; Widding-Havneraas et al., 2021). In the ESORT study, the IV was the hospital’s tendency to operate (TTO), which reflects practice variation across hospitals in ES rates for these conditions (see Appendix C.15.). For each qualifying emergency admission, the TTO was defined as the proportion of eligible emergency admissions in that specific hospital who received ES in the previous 12 mo, thus requiring that the hospital’s past preference for ES strongly predicts treatment choice for the current patient. The rationale for the IV design is that, after adjustment for observed characteristics, the patients’ baseline prognosis is similar across hospitals with different TTO levels (Widding-Havneraas et al., 2021). Hence, the patients can be “randomized” between the ES and NES strategies according to the hospital’s TTO.

While Keele et al. (2018). validated this IV within US claims data, we carefully considered whether each of the above underlying assumptions were met within the EHR data for the ESORT study. We assessed the relevance of the hospital’s TTO with a weak instrument test that is robust to heteroscedasticity and clustering (Olea and Pflueger, 2013). Assumptions (ii), (iii), and (iv) are untestable. The IV would fail the exclusion-restriction condition (assumption iii) if patients admitted to hospitals with high TTO received better care (e.g., postoperative care) leading to lower mortality or shorter stays (and hence costs), regardless of the treatment received, which seems unlikely. However, to increase the plausibility of assumptions (ii) and (iii), we adjusted for a rich set of potential confounders, including proxies for the quality of acute care in each hospital (see Appendix C.5). We assessed the extent to which observed prognostic covariates differed across levels of the instrument (see Figure 4.1). Imbalances observed in measured covariates across levels of the TTO would raise concerns about assumptions (ii) and (iii). We also observed a strong positive, linear relationship between the hospital-level TTO and receipt of ES for all 3 conditions, providing support for assumption (iv).

4.2.2.3 Statistical and sensitivity analyses

LIV estimated PeT effects of ES versus NES on costs and QALYs for each individual allowing for treatment effect heterogeneity and confounding (Armitage and Van Der Meulen, 2010; ESORT Study Group, 2020b; Hernán and Robins, 2016; Jauhari et al., 2020). These were aggregated to report the effects of ES overall and for each prespecified subgroup of interest. Probit regression models were used to estimate the initial propensity score (first stage), whereas GLMs were applied to the cost and QALY data, with the most appropriate chosen according to rMSE (see Appendix C.12). Hosmer-Lemeshow and Pregibon tests were also used to check the model fit and appropriateness (Hosmer and Lemeshow, 2000; Pregibon, 1980). For the QALY endpoint, the logit link and binomial family were selected (all 3 conditions) and, for costs, the log link and Gaussian family (appendicitis and diverticular disease) and the identity link and gaussian family (hernia). Models at both stages adjusted for the above baseline measures, time period, and proxies for hospital quality, defined by rates of emergency readmission and mortality in 2009 to 2010 (time constant) and in the year prior to the specific admission concerned (time-varying; see Appendix C.5).

Figure 4.1. Mean level of rescaled baseline covariates according to the level of the instrumental variable



Overall estimates of incremental costs, QALYs, and INB were reported with standard errors and confidence intervals (CIs) obtained with the nonparametric bootstrap (300 replications), allowing for the clustering of individuals within hospitals and the correlation of individual-level costs and effects. The individual-level estimates of incremental costs and QALYs were also plotted on the cost-effectiveness plane, stratified by subgroups of policy relevance.

The 2SRI and risk-adjustment (GLM regression) approaches took the same approach to model specification and selection (including covariates used for confounding adjustment) to report overall estimates of incremental costs and QALYs and INB. The proportion of missing data across the 3 cohorts was low, with less than 5% missing values for all baseline covariates, other than ethnicity (10% in the appendicitis cohort); thus, a complete case analysis was performed.

Sensitivity analyses

Sensitivity analyses were undertaken to assess whether the results from the main analysis were robust to alternative definitions and assumptions. First, the study adjusted for "quality of care" using external hospital performance measures from the National Emergency Laparotomy Audit (NELA) (NELA Project Team, 2016, 2017, 2018). Second, we considered the sensitivity of our findings to the potential for under- or overestimating costs from EHR data by increasing all costs by 10% (SA2) and to reducing them by 10% (SA3). Third, we considered an alternative approach to QALY calculation that used linear interpolation between the baseline admission, and 1-y follow-up (SA4). Fourth, we considered a longer time horizon of 5 y, by restricting the sample to those patients who were admitted from 2010 to 2014 (SA5).

Ethics approval

The research was approved by the London School of Hygiene and Tropical Medicine ethics committee (Ethics Reference no: 21687). The study involved the secondary analyses of existing pseudo anonymised data and did not require UK National Ethics Committee approval.

4.2.3 Results

The study included 268,144 (appendicitis), 138,869 (diverticular disease), and 106,432 (hernia) patients. The proportions of patients who had ES were 92.3% (appendicitis), 11.4% (diverticular disease), and 58.8% (hernia). The patients with acute appendicitis who had ES were on average younger and more likely to be fit and without comorbidities as compared with those who had NES strategies. For patients with diverticular disease, patients who had ES were less likely to be fit but were of similar age and comorbidity profile to those in the NES groups. For patients with hernia, a higher proportion of women had ES. Other baseline characteristics were similar between the comparison groups (Table 4.1).

The most prevalent forms of ES are listed in Appendix C.13. Most patients in the NES strategy groups did not have an operative procedure.

Table 4.2 presents the unadjusted costs of ES and NES. For patients with diverticular disease, the average total costs for the ES group at 1 y were higher than for the NES group (£16,498 v. £4673), reflecting the higher initial admission costs, including operative costs. For the other 2 conditions, the average 1-y costs of ES versus NES were similar, with the higher operative costs of ES offset by higher readmission costs following the NES strategy (see Appendix C.13). For patients with diverticular disease, before any case-mix adjustment, the proportion of patients who had died by 1 y was higher in the ES versus NES group (see Appendix C.16).

4.2.3.1 IV diagnostics

The hospital's TTO was strongly correlated with ES receipt for all 3 conditions, after case-mix adjustment (see Table 4.3). For the 3 conditions, the F statistic ranged from 135 (appendicitis) to 735 (hernia) versus the commonly applied threshold of 10 (Staiger and Stock, 1997). Thus, the hospital's past preference for ES strongly predicts treatment choice for the current patient. The mean levels of the baseline covariates (rescaled) were similar across the TTO levels (Figure 4.1), which makes it more plausible that the IV also balances unobserved covariates.

Table 4.1. Baseline characteristics of patients in the cohorts

	Acute appendicitis (n=268,144)		**Diverticular disease (n= 138,869)**		**Abdominal wall hernia (n=106,432)**	
	ES (n=247,506)	NES (n=20,638)	ES (n= 15,772)	NES (n=123,097)	ES (n=62,559)	NES (n=43,873)
Gender: n (%)						
Male	134,270 (54)	10,409 (50)	7,074 (45)	49,922 (41)	37,522 (60)	31,341 (71)
Female	113,224 (46)	10,228 (50)	8,698 (55)	73,172 (59)	25,035 (40)	12,530 (29)
Age: mean	38	47	64	64	63	62
SCARF index: n (%)						
Fit	206,796 (84)	15,015 (73)	6,197 (39)	65,911 (54)	33,014 (53)	23,871 (54)
Mild frailty	34,544 (14)	4,052 (20)	5,631 (36)	38,851 (32)	19,608 (31)	13,104 (29)
Moderate frailty	5,041 (2)	1,155 (6)	2,706 (17)	13,433 (11)	7,360 (12)	4,987 (11)
Severe frailty	1,125 (0)	416 (2)	1,238 (8)	4,902 (4)	2,577 (4)	1,911 (4)
Charlson index: n (%)						
0 – comorbidities	207,525 (84)	15,321 (74)	9,789 (62)	73,457 (60)	39,216 (63)	26,297 (60)
1	35,721 (14)	3,989 (19)	4,482 (28)	35,106 (29)	17,494 (28)	12,163 (28)
2	3,715 (2)	1,035 (5)	1,222 (8)	11,454 (9)	4,792 (8)	4,169 (10)
3+ – comorbidities	545 (0)	293 (1)	279 (2)	3,080 (3)	1,057 (2)	1,244 (3)

ES: Emergency surgery, IMD: Index of multiple deprivation, NES: non-emergency surgery, SCARF: secondary care administrative records frailty

Table 4.2. Unadjusted costs of ES and NES strategies (£GBP 2019/20)

	Acute appendicitis (N=268,144)		**Diverticular disease (N=138,869)**		**Abdominal Wall Hernia (N=106,432)**	
	ES (N=247,506)	NES (N=20,638)	ES (N=15,772)	NES (N=123,097)	ES (N=62,559)	NES (N=43,873)
Index admission						
Bed-day costs (£): mean (SD)	1,613 (2,080)	1,850 (3,147)	10,637 (12,919)	1,880 (2,511)	2,249 (7,036)	1,181 (3,853)
Cost diagnostic procedures (£): mean (SD)	28.0 (54.2)	57.8 (69.1)	108 (104)	86.5 (81.4)	20.3 (52.3)	18.2 (45.1)
Cost operative procedures (£): mean (SD)	1,132 (127)	192 (429)	1,947 (938)	1.68 (32.8)	809 (244)	42.3 (209)
Total costs index admission (£): mean (SD)	2,774 (1,974)	2,101 (3,213)	12,690 (13,124)	1,967 (2,537)	3,079 (7,066)	1,242 (3,938)
Readmissions up to 1 year						
Patients with 1+ readmissions: n (%)	66,446 (26.8)	10,895 (53.0)	10,100 (64.2)	90,300 (74.4)	25,947 (41.5)	31,997 (72.9)
Bed-day costs (£): mean (SD)	541 (2,594)	1,408 (4,208)	3,444 (8,028)	2422 (6,167)	1,786 (5,998)	2,581 (7,413)
Cost diagnostic procedures (£): mean (SD)	22.5 (80.2)	70.2 (142)	94.4 (149)	146 (174)	33.5 (100)	45.7 (120)
Cost operative procedures (£): mean (SD)	18.5 (139)	178 (419)	270 (628)	137 (496)	62.7 (242)	406 (457)
Total costs readmissions: mean (SD)	582 (2,650)	1,656 (4,338)	3,808 (6,374)	2,706 (6,743)	1,882 (6,061)	3,033 (7,468)
Total costs at one year: mean (SD)	3,355 (3,519)	3,757 (5,658)	16,498 (16,027)	4,673 (7,145)	4,961 (9,666)	4,275 (8,680)

ES: emergency surgery, NES: non-emergency surgery, SD: standard deviation.

Table 4.3. Instrumental Variable strength for the hospital-level tendency-to-operate (TTO) within the HES data (2009-19) for emergency admissions that met the ESORT study inclusion criteria for each of the three conditions

Condition	Montiel-Pflueger robust weak instrument test F-Statistic
Acute appendicitis	135
Diverticular disease	206
Abdominal wall hernia	735

4.2.3.2 Overall cost-effectiveness results by method

Table 4.4 reports the estimated incremental costs and QALYs and the INB according to the intention-to-treat principle for the overall population using regression adjustment, 2SRI, and the LIV approach. For patients with appendicitis and hernia, all 3 methods reported mean INBs close to zero. For patients with diverticular disease, the results differed by method. The regression adjustment and the 2SRI approaches reported that ES has positive incremental costs, negative incremental QALYs, and negative INBs with 95% CIs below zero (Table 4.4). By contrast, the LIV results show that there was considerable uncertainty in the overall cost-effectiveness estimates for all 3 conditions, with 95% CIs around the INBs that included zero (Table 4.4). For acute appendicitis, the incremental QALYs and costs were also close to zero (Table 4.4). For patients with diverticular disease, the LIV approach reported that, on average, ES led to a cost reduction (−£1724), QALY gain (0.047), and a positive INB (£2664). For patients with abdominal wall hernia, the LIV approach reported that the positive incremental costs of ES (£891) were offset by moderate QALY gains (0.0386; see Appendix C.17).

4.2.3.3 Subgroup analysis of cost-effectiveness of ES

Figure 4.2 reports that beneath the overall LIV results, there is underlying heterogeneity in the INB estimates according to subgroup. For patients with acute appendicitis, ES appears less cost-effective for women, older patients, and those with 2 or 3 comorbidities. For each condition, ES is less cost-effective on average, according to increasing frailty levels. For example, for appendicitis, the estimated INBs for patients with moderate and severe frailty were −£5750 (−£7810, −£3692) and −£18,723 (−£23,886, −£13,561) versus £369 (−£728, £1467) for patients who were fit (see also Appendix C.17).

Table 4.4. Estimated incremental net monetary benefit (INB), costs, and QALYs of ES vs NES strategies

	Acute appendicitis (N=268,144)	Diverticular disease (N=138,869)	Abdominal Wall Hernia (N=106,432)
		INB	
Unadjusted differences	1,431 (1,259, 1,603)	-13,088 (-13,509, -12668)	-303 (-469, -137)
GLM	-165 (-287, -42)	-12,381 (-12,848, -12,058)	-50.1 (-241, 141)
GLM-2SRI	281 (-743, 1,306)	-7,496 (-12,230, -2,763)	-1,474 (-3,038, 2,995)
LIV	-86.2 (-1,163, 991)	2,664 (-4,298, 9,626)	-119 (-1,282, 1,043)
		Incremental costs	
Unadjusted differences	-413 (-513, -312)	11,857 (11,486, 12,228)	674 (548, 800)
GLM	318 (213, 424)	11,266 (10,905, 11,626)	483 (318, 649)
GLM-2SRI	762 (-73.5, 1,598)	5,990 (1,371, 10,609)	1,645 (295, 2,995)
LIV	-109 (-1,130, 913)	-1,724 (-7,878, 4,430)	891 (20.7, 1,762)
		Incremental QALYs	
Unadjusted differences	0.0509 (0.0462, 0.0556)	-0.0616 (-0.0672, -0.0559)	0.0186 (0.0150, 0.0221)
GLM	0.00767 (0.00550, 0.00983)	-0.0594 (-0.0653, -0.0534)	0.0216 (0.018, 0.0253)
GLM-2SRI	0.0522 (0.0294, 0.0750)	-0.0753 (-0.116, -0.0343)	0.0085 (-0.0240, 0.0411)
LIV	-0.00973 (-0.0226, 0.00316)	0.0471 (-0.0829, 0.177)	0.0386 (0.00430, 0.0729)

2SRI: two-stage residual inclusion, GLM: generalised linear model, LIV: local instrumental variable, QALYs: quality-adjusted life years.

Figure 4.2. Estimated Incremental Net monetary Benefit (INB) of ES versus NES strategies for acute appendicitis (panel A), diverticular disease (B) and abdominal wall hernia (C)

(A): Acute appendicitis

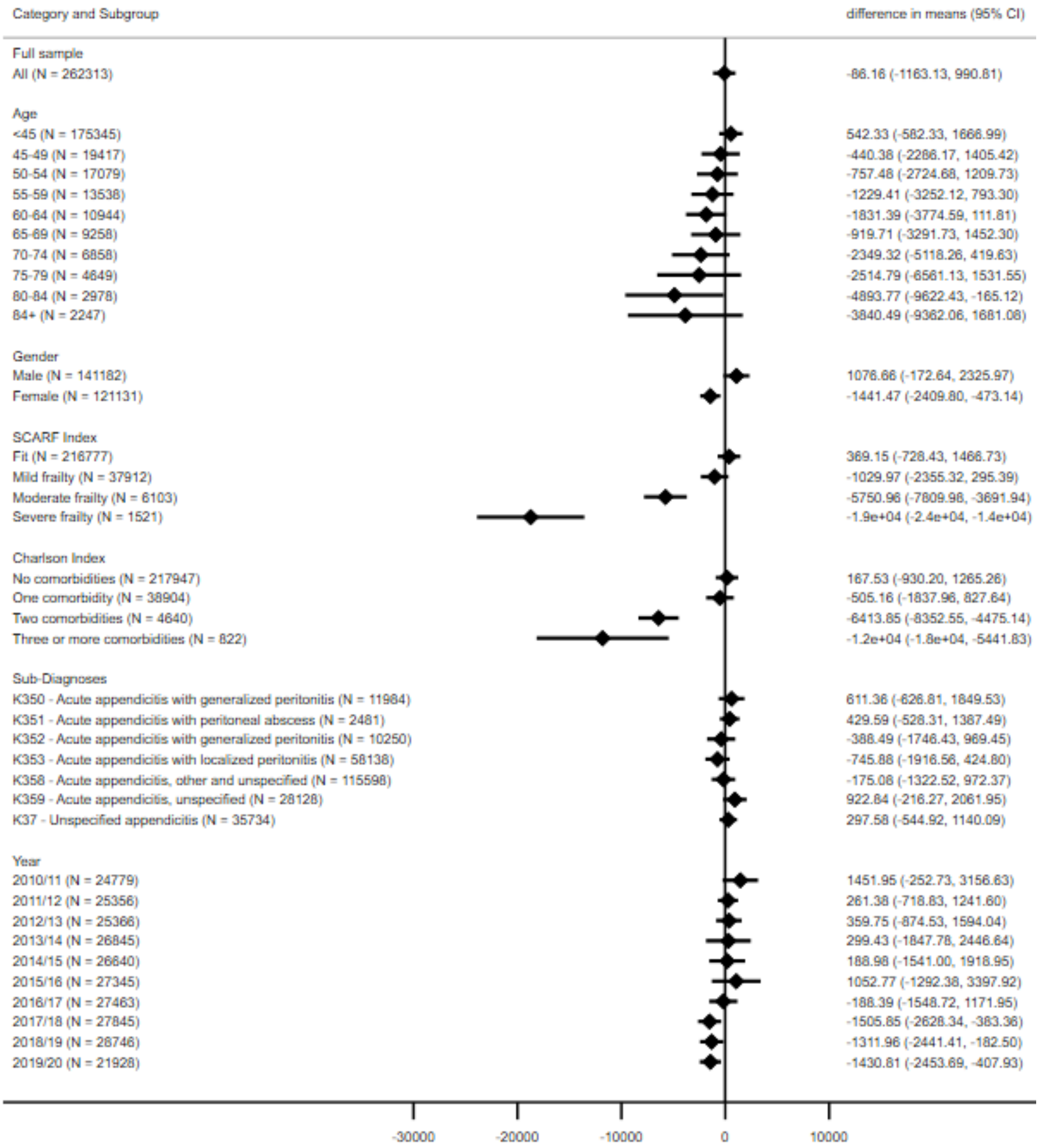


Figure 4.2. (cont.) Estimated Incremental Net monetary Benefit (INB) of ES versus NES strategies for acute appendicitis (panel A), diverticular disease (B) and abdominal wall hernia (C)

(B): Diverticular disease

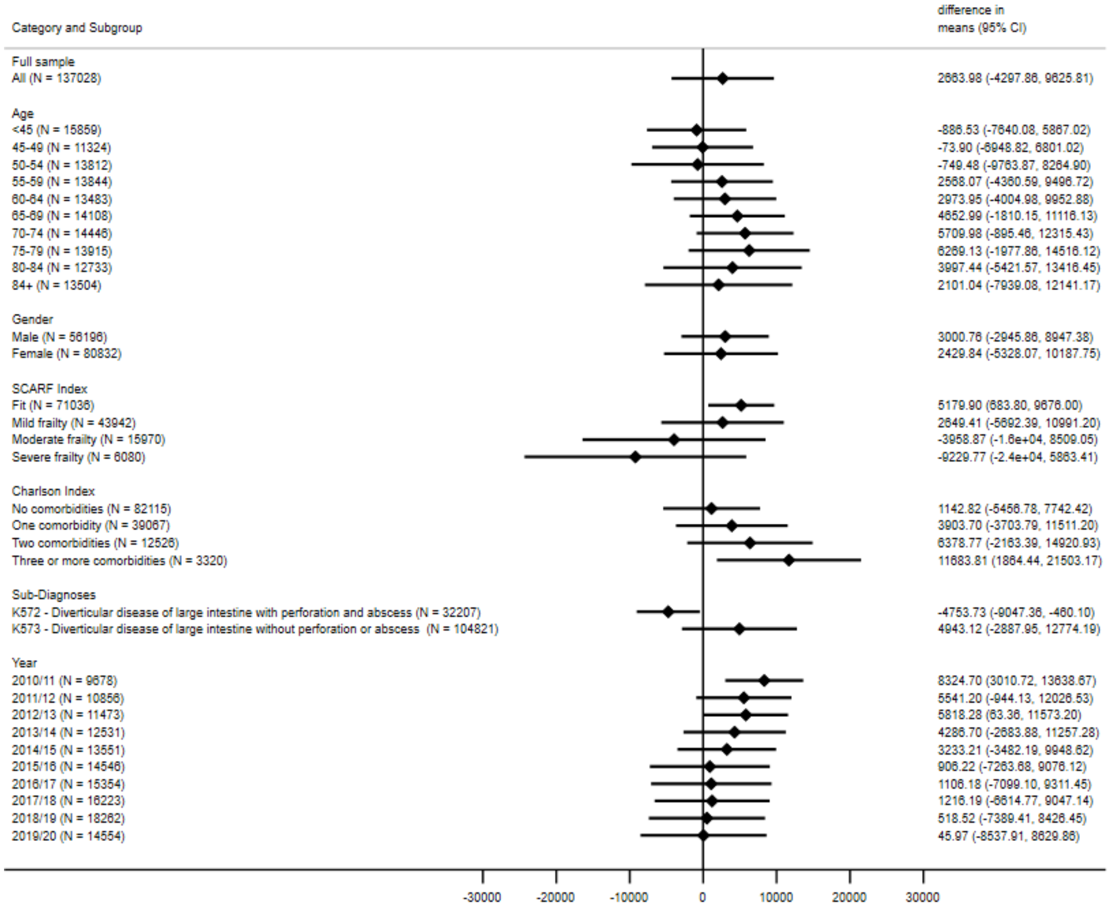
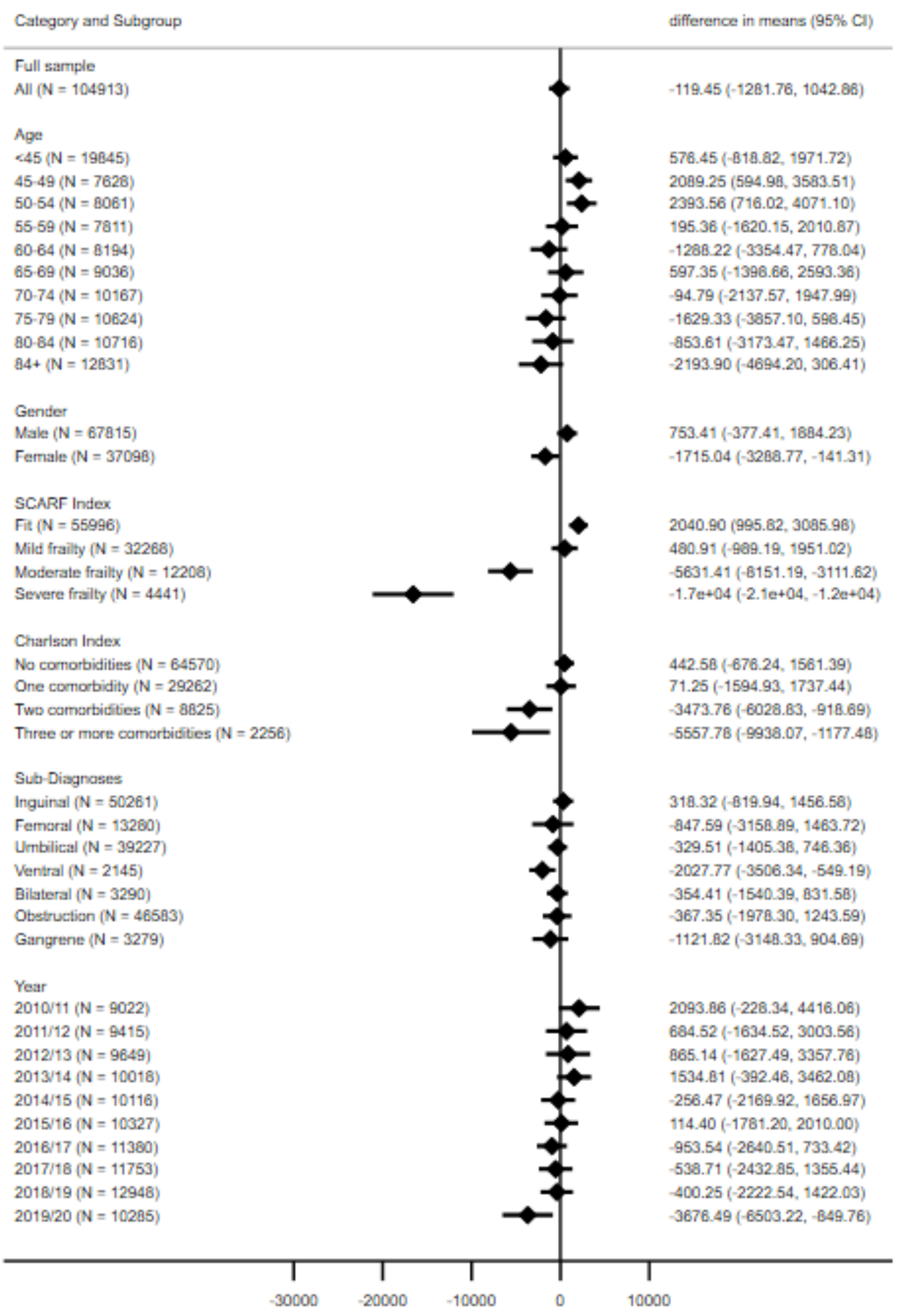


Figure 4.2. (cont.) Estimated Incremental Net monetary Benefit (INB) of ES versus NES strategies for acute appendicitis (panel A), diverticular disease (B) and abdominal wall hernia (C)

(C): Abdominal wall hernia



4.2.3.4 Estimated Individual-level effects of ES on costs and outcomes

Figure 4.3 reports the individual-level estimates of incremental costs and QALYs for the 3 conditions. Here, for illustration, the results are stratified by frailty level. For those with severe frailty, the proportion of patients for whom ES is estimated to be cost-effective is 0.0657% (appendicitis), 46.9% (diverticular disease), and 0.00% (hernia), whereas for patients who were fit, the corresponding proportions were 59.0% (appendicitis), 87.1% (diverticular disease), and 82.0% (hernia).

Figure 4.3. Cost-effectiveness plane of person-centered treatment (PeT) effects on costs and QALYs for appendicitis (panel A), diverticular disease (B) and abdominal wall hernia (C)

(A): Acute appendicitis

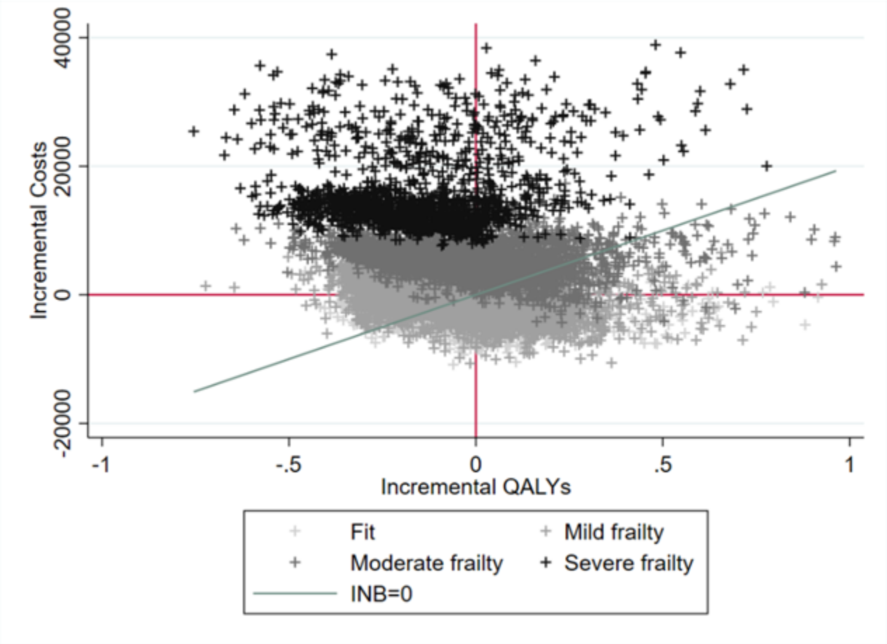
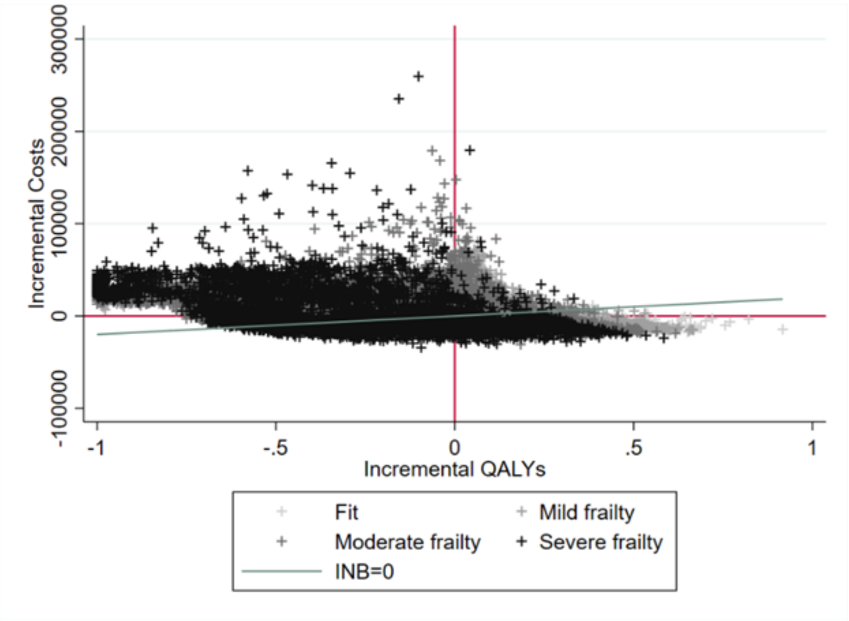
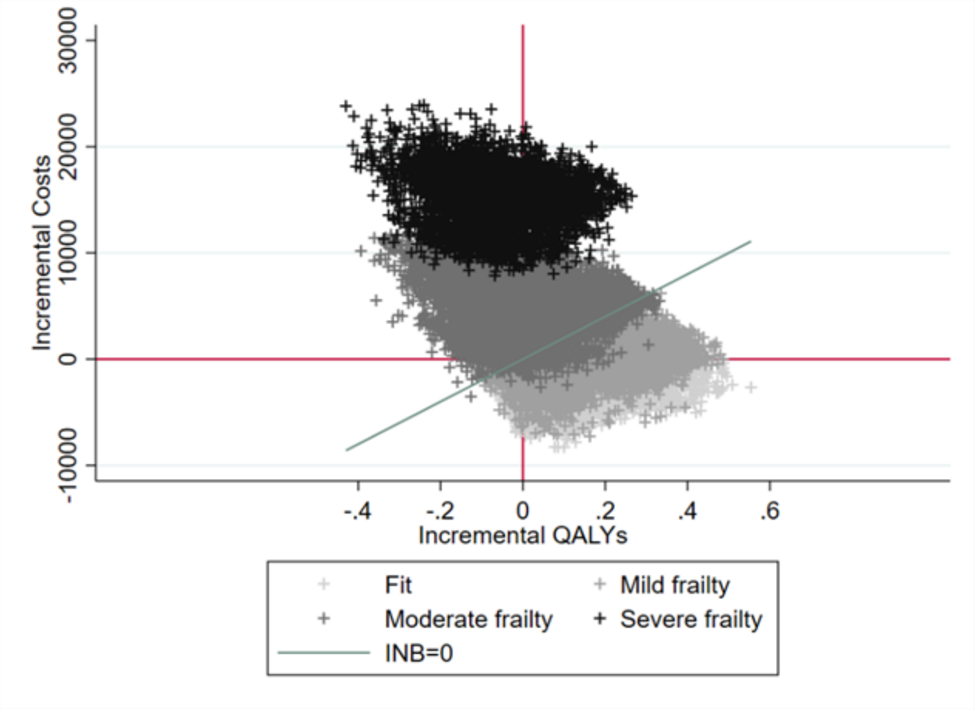


Figure 4.3. (cont.) Cost-effectiveness plane of person-centered treatment (PeT) effects on costs and QALYs for appendicitis (A), diverticular disease (B) and abdominal wall hernia (C)

(B): diverticular disease



(C): abdominal wall hernia



Legend Figure 3: PeT effects of ES on costs and QALYs for appendicitis, diverticular disease and abdominal wall hernia, where each data point relates to one patient in the dataset and each colour to one band of the secondary care administrative records frailty (SCARF) index (fit is light grey, severe frailty is black).

4.2.4 Sensitivity analyses

The overall results were robust to alternative assumptions (see Appendix C.14), including alternative definitions of hospital quality of care (SA1), higher (SA2) or lower (SA3) unit costs, and the use of linear interpolation for calculating QALYs (SA4). The extension to a 5-y time horizon resulted in a negative INB for appendicitis and diverticular disease (SA5), but the sample size was much reduced (~50%), and the CIs surrounding the INB estimates over this extended time horizon were wide and, like the base case, included zero.

4.2.5 Discussion

This article critically examines LIV methods for comparative effectiveness research using EHRs in the context of a CEA. We evaluate the cost-effectiveness of ES compared with NES alternatives for emergency admissions with common acute conditions. The IV design exploited the wide variations in ES rates across hospitals. The LIV method was chosen because it can address confounding and treatment effect heterogeneity, and provide cost-effectiveness estimates for the overall population as well as subpopulations of decision-making relevance, provided the models for the outcome and the treatment assignment are correctly specified. For diverticular disease, the results differed by method. Whereas the traditional approaches reported that, overall, ES was not cost-effective, the LIV approach reported that the overall results were highly uncertain. For appendicitis and hernia, all 3 approaches reported that the overall cost-effectiveness results were uncertain. For all 3 conditions, the LIV approach found heterogeneity in the cost-effectiveness estimates; in particular, ES was not cost-effective for patients with severe levels of frailty.

This article makes 3 important contributions to the literature. First, we add to the literature using IV methods for the evaluation of routinely provided interventions (Basu et al., 2007; Brookhart and Schneeweiss, 2007; Davies et al., 2013; O'Malley et al., 2011; Polsky and Basu, 2012). In the EHR context, given that data are not collected for research purposes, finding a valid IV is especially challenging. This article exemplifies the use of EHRs to substantiate and assess the underlying assumptions of an IV design. For example, to address potential violations of the exclusion restriction, we examined whether the hospital's TTO could minimize imbalances in measured covariates with balance plots and used "internal" (i.e., EHR data) and "external" (i.e., NELA (NELA, 2016, 2017, 2018) information to adjust for the quality of acute care, and improve the plausibility of the exclusion restriction.

Second, this article constitutes a novel application of LIV to a CEA that uses EHR data. We show how EHRs can offer large sample sizes, enabling a CEA to provide precise cost-effectiveness results at the subgroup level, and to reflect the range of patients presenting in routine practice. This article also highlights major challenges of using EHR data for CEA, namely, unmeasured confounding and treatment effect heterogeneity. Although both IV methods considered rely on parametric assumptions and the validity of IV assumptions to address confounding, 2SRI can also fail to identify the ATE in the presence of essential heterogeneity (Chapman and Brooks, 2016; Evans and Basu, 2011). Hence, one interpretation of the differences between the estimates from 2SRI and LIV for patients with diverticular disease is that the estimated effects may differ between marginal patients and the overall population (Chapman and Brooks, 2016). For patients with diverticular disease, patients may well have been selected to receive ES according to measures that were not available in these EHR data, such as the severity of the disease, and so the 2SRI approach may have failed to validly identify the ATE.

Third, this article contributes to the limited previous literature evaluating the cost-effectiveness of ES for these common acute conditions. Some previous studies have also suggested that NES strategies can result in similar outcomes and costs for patients with appendicitis (Flum et al., 2020; Javanmard-Emamghissi et al., 2021; Sippola et al., 2020), whereas others have found NES to be more cost-effective than ES (O'Leary et al., 2021). Published RCTs evaluating ES strategies for acute diverticular disease have failed to recruit sufficiently large populations to explore heterogeneity across population subgroups (Thornell et al., 2016), and are nonexistent for acute hernia. Unlike previous studies (Azhar et al., 2021; Fitzgibbons et al., 2006; Flum et al., 2020; Javanmard-Emamghissi et al., 2021; O'Dwyer et al., 2006; O'Leary et al., 2021; Patel et al., 2020; Salminen et al., 2015; Stroupe et al., 2006; Thornell et al., 2016; Van De Wall et al., 2010; You et al., 2018), the ESORT study included large sample sizes (>100,000 for each condition) and subgroups (e.g., those with severe frailty) excluded from RCTs. These results can help decision makers identify subgroups for whom NES strategies are relatively cost-effective (e.g., patients with severe frailty), those for whom ES is more cost-effective (e.g., "fit" patients), and those for whom there is residual uncertainty and for whom further research may be most valuable (Basu and Meltzer, 2007; Espinoza et al., 2014).

This study has several strengths. First, the study extended a previously validated IV approach, by using large-scale EHR data (Keele et al., 2018). Second, the HES data, while having common features of EHR data (notably the potential for confounding and heterogeneity), were of generally high quality with baseline covariates, all-cause mortality, and resource use data available for ~95% of patients. Third, the study considered 3 different conditions for which it was anticipated there would be heterogeneous treatment effects according to patient subgroups.

While we address some of the challenges of using EHRs for CEA, others remain. First, HRQoL data were not available from HES and had to be obtained from the literature. Granular baseline measures of disease severity (e.g., size of abscess) were not available to provide more nuanced subgroup definitions. Second, it is possible that coding errors within the HES data were incorporated into the estimates of cost and cost-effectiveness, although previous research found that costs estimated from HES data were very similar to those derived from medical records (Thorn et al., 2016). Third, in common with any approach to address confounding, the implementation of the LIV methods made assumptions, in particular, that the relationships of the covariates and the IV, with both the treatment receipt and the outcomes, were correctly specified. Here, more flexible data-adaptive approaches may be helpful, although they have not yet been extended to this context. A further consideration is that subgroup analyses presented here represent the average estimated effect for individuals within the group rather than the causal effect of group membership per se. While the subgroups used here were prespecified within a statistical analysis plan, in other contexts spurious subgroup effects may be obtained by "*P*-hacking."

This article identifies areas for future research. First, future research could build on this work by incorporating data-adaptive methods such as generalized random forests or lasso into the LIV estimation, or by using methods such as causal rule ensembles for exploring heterogeneity (Lee et al., 2020), while recognizing interactions among prognostic variables. Second, the methods used in this study could be extended to chronic diseases by considering other preference-based instruments (e.g., tendency to prescribe), or multiple IV such as genetic markers, which will raise new issues for the LIV approach. Finally, our results can be used to target future trials. For instance, for patients with abdominal wall hernia, there appears to be equipoise about the choice of strategy (~50% in each comparison group). A future trial could collect granular information on patient subgroups, longitudinal HRQoL measures, and be nested within the EHR data to help ensure the results are directly applicable to clinical decision making.

Acknowledgements

We wish to acknowledge the ESORT team for their support throughout this study, including Mr. Paul Charlton (NIHR patient ambassador), Professor Geoff Bellingan, Professor Ramani Moonesinghe, and Beth Silver (project manager). We also wish to acknowledge the ESORT study's independent advisory group for helpful discussions: Mr. Iain Anderson (chair), Professor Sir Nick Black, Mr. Paul Charlton, Dr. Nils Gutacker, Professor Catherine Hewitt, Ms. Susan Moug, and Mr. Ravi Vohra. We would like to thank colleagues who made time, during the early months of the COVID-19 pandemic, to attend the clinical panels. They are Matthew Bedford, Natalie Blencowe, Andrew de Beaux, Martyn Evans, Deepak Hariharan, Deena Harji, Matt Lee, Sonia Lockwood, Frank McDermott, Susan Moug, Dale Vimalachandran, and Ravinder Vohra. Further details are available in the clinical panel report: *https://www.lshtm.ac.uk/media/39151*. Early versions of this article were presented at the summer 2021 Health Economists' Study Group (HESG) meeting hosted by the University of Cambridge and the 2021 international Health Economics Association (iHEA) congress. We would like to acknowledge the feedback from those who participated in both sessions and particularly the paper discussant at HESG, Dr. Caroline Clarke, and at iHEA Professor Andrew Jones. Finally, we thank Professor Anirban Basu for discussions concerning the instrumental variable approaches considered in this article.

References

Abbott TEF, Fowler AJ, Dobbs TD, et al. (2017) Frequency of surgical treatment and related hospital procedures in the UK: A national ecological study using hospital episode statistics. *British Journal of Anaesthesia* 119(2): 249–257. DOI: 10.1093/bja/aex137.

Abercrombie J (2017) *Getting it Right First Time (GiRFT) report. General Surgery.* Available at: http://gettingitrightfirsttime.co.uk/national-general-surgery-report-published-2/.

Angrist J, Imbens G and Rubin D (1993) Identification of causal effects using instrumental variables. *Journal of the American Statistical Association* 91(434): 444–455.

Angrist JD and Imbens GW (1995) Two-Stage Least Squares Estimation of Average Causal Effects in Models with Variable Treatment Intensity. *Journal of the American Statistical Association* 90(430): 431. DOI: 10.2307/2291054.

Angrist JD and Krueger A. (1999) *Empirical Strategies in Labor Economics* (O Ashenfelter and D Cardeds). Handbook o. New York: North-Holland.

Ara R and Brazier JE (2010) Populating an economic model with health state utility values: Moving toward better practice. *Value in Health* 13(5). International Society for Pharmacoeconomics and Outcomes Research (ISPOR): 509–518. DOI: 10.1111/j.1524-4733.2010.00700.x.

Ara R, Brazier J and Zouraq IA (2017) The Use of Health State Utility Values in Decision Models. *PharmacoEconomics* 35. Springer International Publishing: 77–88. DOI: 10.1007/s40273-017-0550-0.

Armitage JN and Van Der Meulen JH (2010) Identifying co-morbidity in surgical patients using administrative data with the Royal College of Surgeons Charlson Score. *British Journal of Surgery* 97(5). Br J Surg: 772–781. DOI: 10.1002/bjs.6930.

Azhar N, Johanssen A, Sundström T, et al. (2021) Laparoscopic Lavage vs Primary Resection for Acute Perforated Diverticulitis: Long-term Outcomes From the Scandinavian Diverticulitis (SCANDIV) Randomized Clinical Trial. *JAMA Surgery* 156(2): 121–128. DOI: 10.1001/jamasurg.2020.5618.

Baiocchi M, Cheng J and Small DS (2014) Instrumental variable methods for causal inference. *Statistics in Medicine* 33(13): 2297–2340. DOI: 10.1002/sim.6128.

Basu A (2014) Estimating person-centered treatment (PeT) effects using instrumental variables: an application to evaluating prostate cancer treatments. *JOURNAL OF APPLIED ECONOMETRICS* 29: 671–691. DOI: 10.1002/jae.

Basu A (2015) Person-centered treatment (PeT) effects: Individualized treatment effects using instrumental variables. *The Stata Journal* 15(2): 397–410.

Basu A and Meltzer D (2007) Value of information on preference heterogeneity and individualized care. *Medical Decision Making* 27(2): 112–127. DOI: 10.1177/0272989X06297393.

Basu A, Heckman JJ, Navarro-Lozano S, et al. (2007) Use of instrumental variables in the presence of heterogeneity and self-selection: an application to treatments of breast cancer patients. *Health Economics* 16(2007): 1133–1157. DOI: 10.1002/hec.1291.

Basu A, Coe NB and Chapman CG (2018) 2SLS versus 2SRI: Appropriate methods for rare outcomes and/or rare exposures. *Health Economics* 27(6): 937–955. DOI: 10.1002/hec.3647.

Basu A, Jones AM and Rosa Dias P (2018) Heterogeneity in the impact of type of schooling on adult health and lifestyle. *Journal of Health Economics* 57. Elsevier B.V.: 1–14. DOI: 10.1016/j.jhealeco.2017.10.007.

Brookhart MA and Schneeweiss S (2007) Preference-based instrumental variable methods for the estimation of treatment effects: Assessing validity and interpreting results. *International Journal of Biostatistics* 3(1): 1–19. DOI: 10.2202/1557-4679.1072.

Chapman CG and Brooks JM (2016) Treatment Effect Estimation Using Nonlinear Two-Stage Instrumental Variable Estimators: Another Cautionary Note. *Health Services Research* 51(6). Blackwell Publishing Inc.: 2375–2394. DOI: 10.1111/1475-6773.12463.

Coleman KJ, Fischer H, Arterburn DE, et al. (2020) Effectiveness of gastric bypass versus gastric sleeve for cardiovascular disease: Protocol and baseline results for a comparative effectiveness study. *JMIR Research Protocols* 9(4): 1–11. DOI: 10.2196/14936.

Cornelissen T, Dustmann C, Raute A, et al. (2016) From LATE to MTE: Alternative methods for the evaluation of policy interventions. *Labour Economics* 41. Elsevier B.V.: 47–60. DOI: 10.1016/j.labeco.2016.06.004.

Davies NM, Gunnell D, Thomas KH, et al. (2013) Physicians' prescribing preferences were a potential instrument for patients' actual prescriptions of antidepressants. *Journal of Clinical Epidemiology* 66(12). Elsevier Inc: 1386–1396. DOI: 10.1016/j.jclinepi.2013.06.008.

ESORT Study Group (2020a) Emergency Surgery Or NoT (ESORT) study. Available at: https://www.lshtm.ac.uk/media/38711.

ESORT Study Group (2020b) Emergency Surgery Or NoT (ESORT) study. Available at: https://www.lshtm.ac.uk/media/39151.

Espinoza MA, Manca A, Claxton K, et al. (2014) The Value of Heterogeneity for Cost-Effectiveness Subgroup Analysis: Conceptual Framework and Application. *Medical Decision Making* 34(8): 951–964. DOI: 10.1177/0272989X14538705.

Evans H and Basu A (2011) *Exploring comparative effect heterogeneity with instrumental variables: prehospital intubation and mortality.* Health, Econometrics and Data Group (HEDG) Working Papers, August. HEDG, c/o Department of Economics, University of York. Available at: https://econpapers.repec.org/RePEc:yor:hectdg:11/26 (accessed 15 June 2021).

Fitzgibbons RJ, Giobbie-Hurder A, Gibbs J., et al. (2006) Watchful waiting vs repair of inguinal hernia in minimally symptomatic men: a randomized clinical trial. *JAMA* 295(3). JAMA: 285–292. DOI: 10.1001/JAMA.295.3.285.

Flum DR, Davidson GH, Monsell SE, et al. (2020) A Randomized Trial Comparing Antibiotics with Appendectomy for Appendicitis. *New England Journal of Medicine* 383(20): 1907–1919. DOI: 10.1056/nejmoa2014320.

Grieve R, O'Neill S, Basu A, et al. (2019) Analysis of Benefit of Intensive Care Unit Transfer for Deteriorating Ward Patients: A Patient-Centered Approach to Clinical Evaluation. *JAMA network Open* 2(2). NLM (Medline): 1–13. DOI: 10.1001/jamanetworkopen.2018.7704.

Heckman JJ and Vytlacil E (2005) Structural equations, treatment effects and econometric policy evaluation. *Econometrica* 73(3): 669–738. DOI: 10.1111/j.1468-0262.2005.00594.x.

Heckman JJ and Vytlacil EJ (1999) Local instrumental variables and latent variable models for identifying and bounding treatment effects. *Proceedings of the National Academy of Sciences of the United States of America* 96: 4730–4734. DOI: 10.1073/pnas.96.8.4730.

Heckman JJ and Vytlacil EJ (2001) Policy-Relevant Treatment Effects. *American Economic Review* 91(2): 107–111. DOI: 10.1257/aer.91.2.107.

Heckman JJ, Urzua S and Vytlacil E (2006) *Understanding instrumental variables in models with essential heterogeneity. NBER Working Paper No. 12574*. Cambridge, MA. DOI: 10.1162/rest.88.3.389.

Herbert A, Wijlaars L, Zylbersztejn A, et al. (2017) Data Resource Profile: Hospital Episode Statistics Admitted Patient Care (HES APC). *International Journal of Epidemiology* 46(4): 1093-1093i. DOI: 10.1093/ije/dyx015.

Hernán MA and Robins JM (2016) Using Big Data to Emulate a Target Trial When a Randomized Trial Is Not Available. *American Journal of Epidemiology* 183(8). Oxford University Press: 758–764. DOI: 10.1093/aje/kwv254.

HM Treasury Department (2020) Gross Domestic Product (GDP) deflators: user guide. Available at: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/205904/GDP_Deflators_User_Guide.pdf (accessed 19 August 2021).

Hosmer DW and Lemeshow S (2000) *Applied Logistic Regression.* 2nd ed. Wiley.

Huber M and Wüthrich K (2019) Evaluating local average and quantile treatment effects under endogeneity based on instruments: a review Evaluating local average and quantile treatment effects under endogeneity based on instruments: a review. *Journal of Econometric Methods* 8(1).

Hutchings A, O'Neill S, Lugo-palacios DG, et al. (2022) Effectiveness of emergency surgery for five common acute conditions: an instrumental variable analysis of a national routine database. *Anaesthesia*: In Press.

Imbens GW and Angrist JD (1994) Identification and Estimation of Local Average Treatment Effects. *Econometrica* 62(2). JSTOR: 467. DOI: 10.2307/2951620.

Jauhari Y, Gannon MR, Dodwell D, et al. (2020) Construction of the secondary care administrative records frailty (SCARF) index and validation on older women with operable invasive breast cancer in England and Wales: A cohort study. *BMJ Open* 10(5): 35395. DOI: 10.1136/bmjopen-2019-035395.

Javanmard-Emamghissi H, Hollyman M, Boyd-Carson H, et al. (2021) Antibiotics as first-line alternative to appendicectomy in adult appendicitis: 90-day follow-up from a prospective, multicentre cohort study. *British Journal of Surgery*: 1–9. DOI: 10.1093/bjs/znab287.

Keele L and Small D (2019) Instrumental variables: Don't throw the baby out with the bathwater. *Health Services Research* 54(3). Blackwell Publishing Inc.: 543–546. DOI: 10.1111/1475-6773.13130.

Keele L, Sharoky CE, Sellers MM, et al. (2018) An instrumental variables design for the effect of emergency general surgery. *Epidemiologic Methods* 7(1). Walter de Gruyter GmbH. DOI: 10.1515/em-2017-0012.

Kennedy EH, Lorch S and Small DS (2019) Robust causal inference with continuous instruments using the local instrumental variable curve. *Journal of the Royal Statistical Society. Series B: Statistical Methodology* 81(1): 121–143. DOI: 10.1111/rssb.12300.

Kreif N, Grieve R and Sadique MZ (2013) Statistical methods for cost-effectiveness analyses that use observational data: a critical appraisal tool and review of current practice. *Health Economics* 22: 486–500. DOI: 10.1002/hec.

Kuo AMS, Thavalathil B, Elwyn G, et al. (2018) The Promise of Electronic Health Records to Promote Shared Decision Making: A Narrative Review and a Look Ahead. *Medical Decision Making* 38(8): 1040–1045. DOI: 10.1177/0272989X18796223.

Kyriacou DN and Lewis RJ (2016) Confounding by Indication in Clinical Research. *JAMA* 316(17): 1786–1797. DOI: 10.1001/jama.2016.14486.

Lee K, Bargagli-Stoffi FJ and Dominici F (2020) Causal Rule Ensemble: Interpretable Inference of Heterogeneous Treatment Effects. Available at: http://arxiv.org/abs/2009.09036.

Manca A, Hawkins N and Sculpher M (2005) Estimating mean QALYs in trial-based cost-effectiveness analysis: The importance of controlling for baseline utility. *Health economics* 14: 487–496. DOI: 10.1002/hec.944.

National Emergency Laparotomy Audit (NELA) Project Team (2016) *Second patient report of the National emergency laparotomy audit*. London.

National Emergency Laparotomy Audit (NELA) Project Team (2017) *Third patient report of the National emergency laparotomy audit*. London. Available at: www.nela.org.uk/reports.

National Emergency Laparotomy Audit (NELA) Project Team (2018) *Fourth patient report of the National emergency laparotomy audit*. London.

National Institute for Health and Care (2013) Guide to the methods of technology appraisal. London. Available at: https://www.nice.org.uk/process/pmg9/chapter/foreword.

O'Dwyer PJ, Norrie J, Alani A, et al. (2006) Observation or Operation for Patients With an Asymptomatic Inguinal Hernia A Randomized Clinical Trial. *Annals of Surgery* 244(2). DOI: 10.1097/01.sla.0000217637.69699.ef.

O'Leary DP, Walsh SM, Bolger J, et al. (2021) A Randomized Clinical Trial Evaluating the Efficacy and Quality of Life of Antibiotic-only Treatment of Acute Uncomplicated Appendicitis: Results of the COMMA Trial. *Annals of surgery* 274(2): 240–247. DOI: 10.1097/SLA.0000000000004785.

O'Malley A., Frank R. and Normand S-LT (2011) Estimating cost-offsets of new medications: Use of new antipsychotics and mental health costs for schizophrenia. *Statistics in Medicine* 30: 1971–1988. DOI: 10.1002/sim.4245.

Ogburn EL, Rotnitzky A and Robins JM (2015) Doubly robust estimation of the local average treatment effect curve. *Journal of the Royal Statistical Society. Series B: Statistical Methodology* 77(2): 373–396. DOI: 10.1111/rssb.12078.

Olea JLM and Pflueger C (2013) A Robust Test for Weak Instruments. *Journal of Business & Economic Statistics* 31(3). Taylor & Francis: 358–369. DOI: 10.1080/00401706.2013.806694.

Patel S V., Hendren S, Zaborowski A, et al. (2020) Evidence-based Reviews in Surgery Long-term Outcome of Surgery Versus Conservative Management for Recurrent and Ongoing Complaints After an Episode of Diverticulitis: Five-year Follow-up Results of a Multicenter Randomized Controlled Trial (DIRECT-Trial). *Annals of surgery*. DOI: 10.1097/SLA.0000000000003920.

Polsky D and Basu A (2012) *Chapter 46: Selection Bias in Observational Data.* (AM Jonesed.). he Elgar C. Edward Elgar Publishing.

Pregibon D (1980) Goodness of Link Tests for Generalized Linear Models. *Journal of the Royal Statistical Society. Series C (Applied Statistics),* 29(1). London: Royal Statistical Society: 14–15. DOI: 10.2307/2346405.

Reynolds K, Barton LJ, Basu A, et al. (2021) Comparative Effectiveness of Gastric Bypass and Vertical Sleeve Gastrectomy for Hypertension Remission and Relapse: The ENGAGE CVD Study. *Hypertension* 78(4): 1116–1125. DOI: 10.1161/HYPERTENSIONAHA.120.16934.

Russell LB (2021) Electronic Health Records: The Signal and the Noise. *Medical Decision Making* 41(2): 103–106. DOI: 10.1177/0272989X20985764.

Salminen P, Paajanen H, Rautio T, et al. (2015) Antibiotic therapy vs appendectomy for treatment of uncomplicated acute appendicitis: The APPAC randomized clinical trial. *JAMA* 313(23): 2340–2348. DOI: 10.1001/jama.2015.6154.

Sippola S, Haijanen J, Viinikainen L, et al. (2020) Quality of Life and Patient Satisfaction at 7-Year Follow-up of Antibiotic Therapy vs Appendectomy for Uncomplicated Acute Appendicitis: A Secondary Analysis of a Randomized Clinical Trial. *JAMA Surgery*: 1–7. DOI: 10.1001/jamasurg.2019.6028.

Staiger D and Stock JH (1997) Instrumental Variables Regression with Weak Instruments. *Econometrica* 65(3): 557. DOI: 10.2307/2171753.

Stewart B, Khanduri P, McCord C, et al. (2014) Global disease burden of conditions requiring emergency surgery. *British Journal of Surgery* 101(1): 9–22. DOI: 10.1002/bjs.9329.

Stroupe KT, Manheim LM, Luo P, et al. (2006) Tension-free repair versus watchful waiting for men with asymptomatic or minimally symptomatic inguinal hernias: a cost-effectiveness analysis. *Journal of the American College of Surgeons* 203(4). J Am Coll Surg: 458–468. DOI: 10.1016/J.JAMCOLLSURG.2006.06.010.

Stürmer T, Funk MJ, Poole C, et al. (2011) Nonexperimental Comparative Effectiveness Research Using Linked Healthcare Databases. *Epidemiology* 22(3): 298–301. DOI: 10.1097/EDE.0b013e318212640c.

Terza J V., Basu A and Rathouz PJ (2008) Two-stage residual inclusion estimation: Addressing endogeneity in health econometric modeling. *Journal of Health Economics* 27(3). NIH Public Access: 531–543. DOI: 10.1016/j.jhealeco.2007.09.009.

Thorn JC, Turner EL, Hounsome L, et al. (2016) Validating the use of hospital episode statistics data and comparison of costing methodologies for economic evaluation: An end-of-life case study from the cluster randomised triAl of PSA testing for prostate cancer (CAP). *BMJ Open* 6(4): 1–7. DOI: 10.1136/bmjopen-2016-011063.

Thornell A, Angenete E, Bisgaard T, et al. (2016) Laparoscopic Lavage for Perforated Diverticulitis With Purulent Peritonitis. *Annals of Internal Medicine* 164(3): 137–145. DOI: 10.7326/M15-1210.

Van De Wall BJM, Draaisma WA, Consten ECJ, et al. (2010) Direct trial. Diverticulitis recurrences or continuing symptoms: Operative versus conservative Treatment. A Multicenter randomised clinical trial. *BMC Surgery* 10: 1–6. DOI: 10.1186/1471-2482-10-25.

Widding-Havneraas T, Chaulagain A, Lyhmann I, et al. (2021) Preference-based instrumental variables in health research rely on important and underreported assumptions: a systematic review. *Journal of Clinical Epidemiology*. Elsevier Inc. DOI: 10.1016/j.jclinepi.2021.06.006.

You K, Bendl R, Taut C, et al. (2018) Randomized clinical trial of elective resection versus observation in diverticulitis with extraluminal air or abscess initially managed conservatively. *British Journal of Surgery* 105(8): 971–979. DOI: 10.1002/bjs.10868.

Chapter 5. How does a local Instrumental Variable Method perform across settings with instruments of differing strengths? A simulation study and an evaluation of emergency surgery

5.1 Preamble to research paper 3

In this chapter, I present a simulation study evaluating the performance of the LIV methodology according to varying levels of IV strength grounded in motivating examples from the ESORT study. This paper follows naturally from research paper 2 (Chapter 4), which raised hypotheses about the requirements for LIV in terms of IV strength. This previous paper helped defined the scenarios of interest in the simulation study, in particular, according to different sample sizes and forms of treatment effect heterogeneity.

As discussed in Chapter 4, while LIV has the potential to inform estimates of policy-relevant parameters when applied to RWD, like any other IV method, it relies on assumptions. In particular, the relevance assumption, relates to the strength of the instrument. If the instrument is not sufficiently strong, that is the correlation of the IV with treatment assignment is insufficient, then, conventional IV approaches do not provide unbiased, statistically efficient estimates of treatment effects.

There is an extensive literature studying the implications of weak IVs for inference. Current practice relies on a rule of thumb, in that to be judged sufficiently strong, the first stage F statistic should exceed a threshold of 10 (Staiger and Stock 1997). However, Lee et al. (2021) showed that 2SLS can have low power at conventional levels of the F statistic and suggested that in order to reduce size distortions of the t-ratio to zero, the first-stage F statistic needs to be much larger. Other recent papers evaluating the finite sample properties of IV methods at have been recently published (Keane and Neal 2021; Angrist and Kolesár 2021; Andrews et al. 2019).

However, no studies have evaluated the requirements in terms of IV strength for LIV methods. While Basu (2014) demonstrated the finite-sample properties of LIV, this paper did not consider scenarios with IVs of moderate strength, or whether the requirements for instrument strength differ according to sample size, or form of treatment effect heterogeneity that is present.

Research paper 3 helps to address this gap in the literature in designing a Monte Carlo simulation to test the performance of LIV in settings with different levels of IV strength. I report performance according bias and statistical efficiency, and contrast LIV against 2SLS over scenarios in which the IV strength, sample size and form of heterogeneity is varied. The findings can inform guidance about the design of future IV studies.

My role involved: reviewing the relevant literatures and, I designed and conducted the simulation study and analysed the case study together with my supervisor SON. I led the interpretation of the main findings. I wrote the first draft version of the manuscript, and incorporated comments from co-authors, SON, AB and RG, into the manuscript.

The analysis received ethical approval from the LSHTM Ethics Committee (ID:21776)

The paper was accepted for presentation at the 29th European Workshop on econometrics and health economics, which was held in September 2022, and has been published at Health Econometrics and Data Group (HEDG) database as a working paper. Following this, the paper is currently being considered for publication in *Health Economics.*

The full reference to the working paper is:

Moler-Zapata, Silvia, Richard Grieve, Anirban Basu, and Stephen O'Neill. 2022. "How Does a Local Instrumental Variable Method Perform across Settings with Instruments of Differing Strengths? A Simulation Study and an Evaluation of Emergency Surgery." 22/18. *Health, Econometrics and Data Group (HEDG) Working Papers.* https://ideas.repec.org/p/yor/hectdg/22-18.html.

References

Andrews I, Stock JH and Sun L (2019) Weak Instruments in Instrumental Variables Regression: Theory and Practice. *Annual Review of Economics* 11: 727–753. DOI: 10.1146/annurev-economics-080218-025643.

Angrist J and Kolesár M (2021) One Instrument to Rule Them All: The Bias and Coverage of Just-Id IV. *SSRN Electronic Journal.* DOI: 10.2139/ssrn.3953944.

Basu A (2014) Estimating person-centered treatment (PeT) effects using instrumental variables: an application to evaluating prostate cancer treatments. *Journal of applied econometrics* 29: 671–691. DOI: 10.1002/jae.

Keane M and Neal T (2021) *A Practical Guide to Weak Instruments. UNSW Economics Working Paper No. 2021-05d.*

Lee D, McCrary J, Moreira MJ, et al. (2021) *Valid T-Ratio Inference for IV. National Bureau of Economic Research Working Paper Series (No. w29124).* DOI: 10.2139/ssrn.3901588.

Staiger D and Stock JH (1997) Instrumental Variables Regression with Weak Instruments. *Econometrica* 65(3): 557. DOI: 10.2307/2171753.

London School of Hygiene & Tropical Medicine
Keppel Street, London WC1E 7HT

T: +44 (0)20 7299 4646
F: +44 (0)20 7299 4656
www.lshtm.ac.uk

RESEARCH PAPER COVER SHEET

Please note that a cover sheet must be completed for each research paper included within a thesis.

SECTION A – Student Details

Student ID Number	1903225	**Title**	Ms
First Name(s)	Silvia		
Surname/Family Name	Moler Zapata		
Thesis Title	METHODS TO ADDRESS CONFOUNDING AND HETEROGENEITY IN COST-EFFECTIVENESS ANALYSIS USING REAL-WORLD DATA		
Primary Supervisor	Prof Richard Grieve		

If the Research Paper has previously been published please complete Section B, if not please move to Section C.

SECTION B – Paper already published

Where was the work published?			
When was the work published?			
If the work was published prior to registration for your research degree, give a brief rationale for its inclusion			
Have you retained the copyright for the work?*	**No**	Was the work subject to academic peer review?	**No**

*If yes, please attach evidence of retention. If no, or if the work is being included in its published format, please attach evidence of permission from the copyright holder (publisher or other author) to include this work.

SECTION C – Prepared for publication, but not yet published

Where is the work intended to be published?	Health Economics
Please list the paper's authors in the intended authorship order:	Silvia Moler Zapata, Richard Grieve, Anirban Basu, Stephen O'Neill

Stage of publication	**Not yet submitted**

SECTION D – Multi-authored work

For multi-authored work, give full details of your role in the research included in the paper and in the preparation of the paper. (Attach a further sheet if necessary)	My role involved: reviewing the relevant literatures and, I designed and conducted the simulation study and analysed the case study together with my supervisor SON. I led the interpretation of the main findings. I wrote the first draft version of the manuscript, and incorporated comments from co-authors, SON, AB and RG, into the manuscript.

SECTION E

Student Signature	Silvia Moler Zapata
Date	24 Sept 2022

Supervisor Signature	R Grieve
Date	24 Sept 2022

5.2 Research paper 3: How does a local Instrumental Variable Method perform across settings with instruments of differing strengths? A simulation study and an evaluation of emergency surgery.

Authors

Silvia Moler-Zapata[1], Richard Grieve[1], Anirban Basu[2,3] and Stephen O'Neill[1].

Affiliations:

[1]Department of Health Services Research and Policy, London School of Hygiene & Tropical Medicine, London, UK
[2]The Comparative Health Outcomes, Policy, and Economics (CHOICE) Institute, Department of Pharmacy, and Departments of Health Services and Economics, University of Washington, Seattle, USA.
[3]National Bureau of Economic Research, Cambridge, MA, USA.

Abstract

Local instrumental variable (LIV) approaches use continuous/multi-valued instrumental variables (IV) to generate consistent estimates of average treatment effects (ATEs) and Conditional Average Treatment Effects (CATEs). However, there is little evidence on how LIV approaches perform with different sample sizes or according to the strength of the IV (as measured by the first-stage F-statistic). We examined the performance of an LIV approach and a two-stage least squares (2SLS) approach in settings with different sample sizes and IV strengths, and considered the implications for practice.

Our simulation study considered three sample sizes (n = 5000, 10000, 50000), six levels of IV strength (F-statistic = 10, 25, 50, 100, 500, 1000) under four 'heterogeneity' scenarios: effect homogeneity, overt heterogeneity (over measured covariates), essential heterogeneity (over unmeasured covariates), and overt and essential heterogeneity combined. Compared to 2SLS, the LIV approach provided estimates for ATE and CATE with lower levels of bias and RMSE, irrespective of the sample size or IV strength. With smaller sample sizes, both approaches required IVs with greater strength to ensure low (<5%) levels of bias. In the presence of overt and/or essential heterogeneity, the LIV approach reported estimates with low bias even when the sample size was smaller (n = 5000), provided that the instrument was moderately strong (F-statistic greater than 50, for the ATE estimand).

We considered both methods in evaluating emergency surgery across three different acute conditions with IVs of differing strengths (F-statistic ranging from 100 to 9000), and sample sizes (100000 to 300000). We found that 2SLS did not detect significant differences in effectiveness across subgroups, even with subgroup by treatment interactions included in the model. The LIV approach found there were substantive differences in the effectiveness of emergency surgery according to subgroups; for each of the three acute conditions, frail patients had worse outcomes following emergency surgery.

These findings indicate that when a continuous IV of a moderate strength is available, LIV approaches are better suited than 2SLS to estimate policy-relevant treatment effect parameters.

Keywords

Instrumental Variables, Instrument Strength, Tendency to Operate, Emergency Surgery.

5.2.1 Introduction

The personalisation of treatment choice can be informed by comparative effectiveness research that exploits the widespread availability of electronic health records (EHRs), but requires methods that address confounding and heterogeneity. For conventional linear Instrumental Variable (IV) methods, such as two-stage least squares (2SLS) to identify policy-relevant estimands such as the Average Treatment Effect (ATE) or Conditional Average Treatment Effects (CATEs), it is required that there is no essential heterogeneity (Heckman et al., 2006). Essential heterogeneity arises when treatment effects differ over levels of unmeasured confounders, in which case 2SLS no longer identifies the ATE, even if the instrument is strong and valid (Heckman et al., 2006). Essential heterogeneity, is a major concern in health care, as it is commonly the case that there are biological correlations between risk factors, some of which remain unobserved to the analyst.

In the presence of essential heterogeneity, Local Instrumental Variable (LIV) approaches can provide consistent estimates of the ATE and CATEs (Heckman and Vytlacil, 2005). LIV methods draw on theory about individual's choices to identify 'marginal treatment effects' (MTEs) for individuals at the 'margin of treatment choice' (Bjorklund and Moffitt, 1983; Heckman and Vytlacil, 1999). These MTEs are identified for individuals for whom the level of the IV is such that observed characteristics encouraging treatment (including the IV) and unobserved characteristics discouraging treatment are balanced, so there is equipoise about the treatment decision. Here, a small change (or nudge) in the level of a valid, continuous IV 'tips the balance' for the treatment decision for these marginal patients, without changing the distribution of the underlying risk factors. Therefore, comparing mean outcomes between two groups of patients only separated by a small change in the IV, identifies MTEs for individuals who *comply* with the change in treatment, due to that small change in the IV. A continuous instrument with sufficient support allows all individuals to be defined as 'compliers' at some level of the IV (Heckman and Vytlacil, 1999). Hence, given observed covariates, MTEs can be estimated along the continuum of the IV, and aggregated to provide CATEs and ATEs (Heckman and Vytlacil, 1999, 2001, 2005)

The theoretical properties of these LIV methods in settings with essential heterogeneity have been discussed by Heckman et al. (2006), Basu et al. (2007) and Angrist and Fernández-Val (2011). However, most simulation studies of IV methods only consider treatment effects that are homogeneous, or heterogenous according to measured factors (overt heterogeneity) (Martínez-Camblor et al., 2019; Terza, Basu, et al., 2008; Terza, Bradford, et al., 2008). Studies that have considered essential heterogeneity, have found that 2SLS provides inconsistent estimates of the ATE (Basu, Coe, et al., 2018; Brooks et al., 2018; Chapman and Brooks, 2016), whereas Basu (2014) reports that a LIV method could provide consistent estimates of the ATE and CATE in finite samples. LIV methods have now been applied across a multitude of settings including cardiovascular and bariatric surgery, universal child care programs and transfers to intensive care units (Basu, Jones, et al., 2018; Cornelissen et al., 2018; Grieve et al., 2019; Reynolds et al., 2021).

A major barrier to wider use of IV approaches in general is that if the instrument is only weakly associated with treatment assignment, then IV estimators can provide very biased and imprecise estimates (Bound et al., 1995; Nelson and Startz, 1990; Stock et al., 2002). Weak IVs can also amplify the bias arising due to violations of the other assumptions (Bound et al., 1995; Small and Rosenbaum, 2008). While current practice tends to rely on the first-stage F-statistic exceeding the value of 10, (Staiger and Stock, 1997) recent developments in the weak identification literature for IV models have revealed the shortcomings of an unequivocal decision rule for assessing weak identification (Andrews et al., 2019; Angrist and Kolesár, 2021; Keane and Neal, 2021; Lee et al., 2021; Moffitt and Zahn, 2022). For LIV to provide consistent, precise estimates of ATE or CATEs, requires a strong continuous/multi-valued IV with sufficient support to ensure that there is a level of the IV at which each unit 'complies' (i.e., is selected into treatment according to the level of the IV). However, no study has assessed the levels of IV strength that are required for an LIV estimator to perform well, nor how performance may differ according to the sample size available, in settings with essential heterogeneity.

This paper addresses this gap in the literature by contrasting LIV with the commonly used 2SLS estimator in Monte Carlo simulations, motivated by a case study which highlights typical issues pertaining to heterogeneity, sample size and IV strength. We simulate four scenarios: two of them under restrictive assumptions about heterogeneity (A: homogeneity; B: overt heterogeneity), one where treatment effects are allowed to be heterogenous according to an unmeasured confounder (C: essential heterogeneity), and one where both forms of heterogeneity are present (D: overt and essential heterogeneity). Across all scenarios, ATE and CATE are the parameters of interest.

This paper is structured as follows. In section 5.2.2, we outline the motivating example. In section 5.2.3, we define the estimands and identification assumptions for

2SLS and LIV, and present the methods for the simulation study. In section 5.2.4, we present the results of the simulation study and the case study. In section 5.2.5, we discuss how this study adds to the literature and the implications for further research.

5.2.2 Motivating example: the ESORT study

The ESORT (Emergency Surgery OR noT) study evaluated the effectiveness of emergency surgery for acute gastrointestinal conditions. The primary outcome of the study was the number of 'days alive and out of hospital' (DAOH) at 90-days (see Hutchings et al. (2022) for details), which encompasses mortality and total length of hospital stay (LOS). The study exemplifies the key issues that arise when applying IV methods to EHR data to provide policy-relevant estimates of comparative effectiveness (ESORT Study Group, 2020; Hutchings et al., 2021, 2022). Patients presented as emergency admissions and were selected for either emergency surgery (ES), or alternative interventions such as medical management or delayed surgery, according to unmeasured characteristics such as the severity of the disease, and hence unmeasured confounding and essential heterogeneity were major concerns.

The ESORT study followed Keele et al. (2018) and developed a continuous preference-based IV for ES receipt to evaluate the effectiveness of ES for three acute gastrointestinal conditions: acute appendicitis, gallstone disease and abdominal wall hernia, using routine hospitalisation data from the hospital episode statistics (HES) inpatient database in England. The IV was the hospital's tendency to operate (TTO), a proxy measure of the hospital's latent preference for ES, defined as the proportion of eligible emergency admissions in each of 174 hospitals who had ES in the year preceding each admission. Given a relevant IV, two main assumptions need to hold: (i) conditional on the variables included in the models, the hospital's TTO was not correlated with the patient's outcome except through treatment assignment, (ii) it does not increase the probability of treatment for an individual at some value of the IV, but decrease it for higher values. The study design had some important features to support this assumption. First, in this emergency setting patients were unlikely to select the hospital according to quality of care. Second, the study only included direct admissions to hospital, so there was no scope to transfer the patient according to physician or patient choice. Third, information was collated on a rich set of proxies for the hospital's quality of acute care, including rates of mortality and emergency admissions in previous years, which were included in the models as fixed effects. Fourth, observed covariates, were balanced across all levels of the TTO, which helped support the requisite assumption that the IV also balanced unmeasured confounders

(Hutchings et al., 2022; Moler-Zapata et al., 2022). The requisite assumption that the IV has a monotonic effect on treatment receipt could not be formally tested on the data. However, it was deemed plausible in this setting, as it seems unlikely that there are patients who would receive emergency surgery when admitted to hospitals with low TTO but receive NES when admitted into a hospital with high TTO.

The ESORT study highlighted several outstanding concerns pertaining to IV methods in general, and LIV approach in particular. While the study reported estimates of the ATE, from the outset, there was policy interest in estimating the CATEs, according to baseline covariates including age, number of comorbidities, and levels of frailty. While the sample sizes for each condition, were relatively large, they also differed across conditions, from 268,144 (appendicitis) and 240,977 (gallstone disease), to 106,432 (hernia) patients. There were also differences in the strength of the IV with F-statistics ranging from 141 (acute appendicitis), 739 (hernia) to 9,053 (gallstone disease). Hence, the ESORT study further motivated the interest in what strength of continuous IV was required to provide unbiased, efficient estimates of policy relevant estimands such as CATEs in settings with essential heterogeneity, and according to different sample sizes.

5.2.3 Methods

5.2.3.1 Instrumental variables methods

Throughout we use the Neyman-Rubin potential outcomes framework (Neyman, 1990; Rubin, 1974). Let Y_D denote the observed outcome, D_Z denote the treatment received, and Z denote the instrumental variable, such that we observe (Y_D, D_Z, Z) for each individual. For each patient, let $Y_1 = \mu_1(X_O, X_U, \vartheta)$ and $Y_0 = \mu_0(X_O, X_U, \vartheta)$ denote the potential outcomes, where X_O is the vector of observed covariates, X_U is a vector of unmeasured confounders, and ϑ captures all the remaining unobserved random variables. Throughout, we assume exogeneity of the covariates (A1), so that the treatment assignment is the only source of endogeneity, such that $(X_O, X_U) \perp \vartheta$ and $X_O \perp X_U$.

5.2.3.2 Identification assumptions

Angrist et al. (1993) defined a series of structural assumptions for the identification of the LATE. Here, following Abadie (2003) and Tan (2006) we make the following assumptions which are the conditional version of the assumptions outlined by Angrist et al. (1993):

(A2)	Unconfoundedness of Z	$(Y_{d_z}, D_z) \perp Z \vert X_O$
(A3)	Exclusion restriction	$Y_{d_z} = Y_d$ with probability 1
(A4)	Relevance	$0 < P(Z = z) < 1$
(A5)	Monotonicity	If $z' > z$ then $D_{z'} \geq D_z$ with probability 1
(A6)	Stable Unit Treatment Value Assumption	$D = D_Z$ and $Y = Y_D$

Assumption (A2) requires that Z is as good as randomly assigned within levels of X_O. Assumption (A3) rules out the possibility that Z has a direct effect on the outcome other than through D_z. Assumptions (A2) and (A3) ensure that the only effect of the Z on the outcome is through D_z. This is sometimes called the independence assumption. Assumption (A4) ensures that Z and D_z are correlated conditional on X_O. Assumption (A5) requires that an increase in Z always results in a higher or equal level of treatment assignment. Assumption (A6) requires that one individual's potential outcomes (Y_D) and treatments (D_z) are not influenced by other individuals' levels of Z (i.e., no interference), nor by how the instrument or treatment is delivered (i.e., no different versions of Z or D_z).

5.2.3.3 Estimands

Imbens and Angrist (1994) and Angrist et al. (1993) show that, under the assumptions outlined above, the LATE can be defined as $\Delta^{LATE}(x_o, z, z') = E[Y_1 - Y_0 | X_O = x_o, D_z < D_{z'}]$ and is identified by the IV estimand:.

$$\frac{E[Y|X_O = x_o,\ Z = z'] - E[Y|X_O = x_o,\ Z = z]}{E[D|X_O = x_o,\ Z = z'] - E[D|X_O = x_o,\ Z = z]}$$

Vytlacil (2002) and Tan (2006) showed that the independence (A2 and A3) and monotonicity assumptions (A5) of the LATE framework are equivalent to those imposed by a non-parametric selection model, where treatment assignment depends on whether a latent index ($\mu_D(X_O, Z)$) crosses a particular threshold (X_{U_D}):

$$D_z = 1\{\mu_D(X_O, Z) \geq X_{U_D}\}$$

where X_{U_D} is a random variable that captures X_U and all other factors influencing treatment assignment but not the outcomes. As in Heckman and Vytlacil (1999, 2001) we can rewrite this equation as D_z=1$\{P(X_O, Z) > V\}$, where $V = F_{X_{U_D}}[X_{U_D} | X_O =$

$x_O, Z = z]$ with $V \perp (Z, X_O)$ and $P(x_O, z) = F_{X_{U_D}|x_O,z}[\mu_D(X_O, Z)]$ is the propensity for treatment, and F represents a cumulative distribution function. Therefore, for any arbitrary distribution of X_{U_D} conditional on X_O and Z, by definition $V \sim Uniform[0,1]$ conditional on X_O and Z. Then, the MTE can be defined as, $\Delta^{MTE}(x_O, p) = E(Y_1 - Y_0|X_O = x_O, V = v)$ and Heckman and Vytlacil (1999, 2001) showed that, under the standard IV assumptions, it can be identified by:

$$\frac{\partial E_{\vartheta}(Y|X_O = x_o, Z = z)}{\partial p} = E_{\vartheta}[(Y_1 - Y_0)|X_O = x_o, V = v]$$

MTEs can be aggregated directly to obtain estimates of the ATE as shown in Heckman et al. (2006). Basu (2014) showed that MTEs can be used to derive personalised treatment (PeT) effects for each individual that take into account the plausible range of values that V may take for each patient, in addition to their observed covariates, IV and actual treatment assignment (see Section 5.2.3.2) (Basu, 2014). The rationale for this approach is that the treatment assignment status provides some information on X_{U_D}. For patients in the treatment group ($D_z = 1$), the propensity to choose treatment based on X_O and Z must outweigh the propensity to choose the comparator strategy based on X_{U_D}, i.e., $P(x_O, z) > v$. For patients in the comparator strategy ($D_z = 0$), the opposite is true. The PeT effect for an individual is obtained by averaging the MTEs corresponding to that individual's level of X_O and Z over those values of unobserved variables that are compatible with that patient's treatment assignment. Hence, $\Delta^{PeT}(x_O, p, D) = E(Y_1 - Y_0|X_O = x_O, P(z, x_O) > v)$ for individuals with $D_z = 1$ and $\Delta^{PeT}(x_O, p, D) = E(Y_1 - Y_0|X_O = x_O, P(z, x_O) < v)$ for individuals with $D_z = 0$.

All of the treatment effect estimands, including ATE and CATEs, can be derived by appropriately aggregating the PeT effects since these are defined at the individual level (see section 5.2.3.4)

5.2.3.4 Estimation methods

5.2.3.4.1 Two-stage Least Squares estimator

2SLS is a common approach to the implementation of IV methods that consistently estimates the ATE parameter under homogeneity, or the LATE parameter under essential heterogeneity given a binary IV. Under assumptions (A1)-(A6), the 2SLS (Wald) estimator involves: (i) estimating $E[D_Z|X_O, Z]$ by regressing D_z on X_O and Z, and (ii) estimating $E[Y_D|D_z, X_O, Z]$ by regressing on X_O and $\hat{E}[D_Z|X_O, Z]$. When the instrument is continuous, 2SLS reports a weighted average of LATEs, which requires careful interpretation (Baiocchi et al., 2014).

5.2.3.5 Local Instrumental Variables estimator: estimating PeT effects

Basu (2014, 2015) describe in detail the series of steps required to estimate PeT effects using the LIV methodology. Briefly, D_z is regressed on Z and X_O, as above, using appropriate methods for binary outcomes and the propensity for treatment $p(x_O, z)$ is estimated. Next, Y is regressed on X_O and a function of $\hat{p}(x_O, z)$ including interactions with X_O. The approach outlined in Basu (2014) involves differentiating the outcome model $g(Y)$ by $\hat{p}(x_O, z)$. Next, PeT effects for each individual can be obtained by performing numerical integration, with MTE $(\partial \hat{g}(Y)/\partial \hat{p})$ evaluated by replacing $\hat{p}$ using 1,000 random draws of $u \sim unif(\min(\hat{p}(x_O, z)), \max(\hat{p}(x_O, z)))$. Then, $D^* = \Phi^{-1}\{\hat{p}(x_O, z)\} + \Phi^{-1}(1-u)$ can be computed. PeT effects can be computed by averaging $\partial \hat{g}(Y)/\partial \hat{p}$ over values of u for which $D^* > 0$ if $D = 1$; or over values of $D^* \leq 0$ if $D = 0$. Finally, averaging PeT effects over all of the observations provides an estimate of the ATE for the population, and over strata of X_O gives the CATE for the subpopulation of interest. Standard errors can be computed using bootstrap methods (Basu, 2015). We now consider the design of the simulation study to contrast the relative performance of the LIV and 2SLS approaches.

5.2.4 Simulation study

Motivated by the gaps in the extant literature, and the motivating example, this simulation study was designed to consider the relative performance of 2SLS and LIV approaches across settings that differed with respect to the form of heterogeneity, the sample size and the strength of the IV. We report the performance of the methods in a Monte Carlo Simulation study according to their mean bias (%) and Root Mean Squared Error (RMSE) for each estimand (ATE and CATE).

5.2.4.1 Data Generating process

We create 5,000 datasets each containing *N*= $\{5000,\ 10000,\ 50000\}$ units, of which 50% are assigned to the treated group. The data generating process (DGP) includes one observed (X_O) and one unmeasured (X_U) covariate. We draw X_O, X_U and the instrument, Z from normal distributions with mean 0, and standard deviation 3. Three subgroups of interest are defined by whether the individuals' values for X_O are more than 0.5 standard deviations below or above its mean.

5.2.4.1.1 Treatment model

The treatment assignment is determined by the latent variable D^*, defined as:

$$D^* = \delta_D + 3X_O - 3X_U + \delta_Z Z + (4 - \delta_Z)\epsilon_D$$

where ϵ_D has a normal distribution with mean 0 and standard deviation, 1. Treatment is then determined as $D = 1$ if $D^* > 0$ and $D = 0$ otherwise. The parameters δ_Z and δ_D are chosen to ensure the IV F-statistic, $F_{IV},$ equals the desired level $F_{Target} = \{10, 25, 50, 100, 500, 1000\}$ on average, with,

$$F_{IV} = (N - df_m - 1) * \frac{\sigma^2_{no\ IV} - \sigma^2_{IV}}{\sigma^2_{IV}}$$

where $\sigma^2_{no\ IV}$ and σ^2_{IV} indicate the residual variance from regressing D on X_O with or without including the IV respectively, and df_m is the number of parameters in the model excluding the IV (i.e., $df_m = 2$ here). For a given F-statistic, a larger sample size implies a lower compliance rate, which in turn will imply a weaker instrument. At low compliance rates, the MSE of IV estimates can increase substantially (Little et al., 2009). We estimate the compliance rate for each sample size and F-statistic, by contrasting treatment uptake at the 1st and 99th percentiles of the IV.

5.2.4.1.2 Outcome model

The outcome models under treatments (Y_1) and control (Y_0) can be written as:

$$Y_0 = \beta_0 + \beta_1 X_O + \beta_2 X_U + \epsilon_{Y_0}$$

$$Y_1 = (\beta_0 + \tau_0) + (\beta_1 + \tau_1)X_O + (\beta_2 + \tau_2)X_U + \epsilon_{Y_1}$$

Implying the treatment effect is $\tau = E(Y_1 - Y_0) = \tau_0 + \tau_1 X_O + \tau_2 X_U$. Specifically, we define the outcome under control as follows:

$$Y_0 = -10 - 10X_O + 10X_U + N(0,1)$$

We consider 4 scenarios for the outcome under treatment, Y_1. In Scenario A, effects are homogeneous ($\tau = 50$). In Scenario B, effects are heterogeneous but depend only on observed confounders (overt heterogeneity) ($\tau = 40 + 20X_O$). In Scenario C, X_U influences both the treatment assignment and the gains from treatment ($\tau = 40 + 20X_U$). In this Scenario, there is essential heterogeneity but no overt effect heterogeneity. Finally in Scenario D there is both overt and essential heterogeneity ($\tau = 20 + 20X_O + 20X_U$). Table 5.1 displays the parameter values for each scenario.

The parameter combinations of interest consist of combinations of $n = \{5000, 10000, 50000\}$ and $F_{Target.} = \{10, 25, 50, 100, 500, 1000\}$. For each parameter combination for each scenario, we create 5000 datasets using the DGP described above and estimate the treatment effects as described below.

Table 5.1. Definition of the simulation scenarios

	Sample size	**F-statistic**	$\boldsymbol{\tau_0}$	$\boldsymbol{\tau_1}$	$\boldsymbol{\tau_2}$
Scenario A: Homogeneity	All sample sizes ($n = \{5000, 10000, 50000\}$)	All F-statistic values ($F_{Target} = \{10, 25, 50, 100, 500, 1000\}$)	50	0	0
Scenario B: Overt heterogeneity			40	20	0
Scenario C: Essential heterogeneity			40	0	20
Scenario D: Overt and essential heterogeneity			20	20	20

5.2.4.2 Implementation of methods

For the 2SLS model, we control for X_O and instrument D by Z. To capture heterogeneity, we also include an interaction between X_O with D, and instrument this with interactions of Z and X_O. To obtain effect estimates, we use the recycled predictions approach, whereby the two potential outcomes (Y_0 and Y_1) are predicted from the second stage model after setting $D = 0$ or $D = 1$ and the interaction $X_O{*}D = 0$ or X_O (Basu and Rathouz, 2005; Stata Corp Lp, 2001). The individual level effect is then estimated as $\hat{\tau} = \hat{\mathrm{Y}}_1 - \hat{\mathrm{Y}}_0$, allowing us to calculate the ATE, and CATEs for the three subgroups ($\mathrm{CATE}_{1,}$ $\mathrm{CATE}_{2,}$ and CATE_3).

For the LIV approach, we first estimate the propensity for treatment conditional on $X_O$and Z, and in the second stage outcome model we include X_O, D, the estimated propensity score ($\hat{p}$), $\hat{p}{*}X_O$ and $\hat{p}^{\,2}$. (Heckman and Vytlacil, 2005). We then estimate PeT effects for each individual as described in Basu (2015) using the petiv command in Stata. The estimated PeT effects are then aggregated to obtain estimates of the ATE, $\mathrm{CATE}_{1,}$ $\mathrm{CATE}_{2,}$ and $\mathrm{CATE}_{3.}$ Before applying either method, we remove observations at those levels of the estimated propensity score where there is insufficient overlap (Basu, 2015).

5.2.5 Results

5.2.5.1 Simulation study

Figures 1-4 present mean (%) bias in the ATE and CATE estimates (Figure 5.1 and Figure 5.2, respectively) and the corresponding plots for RMSE (Figure 5.3 and Figure 5.4, respectively). The results for the three subgroups showed similar patterns, and hence, for brevity, we only report the results for one of them.

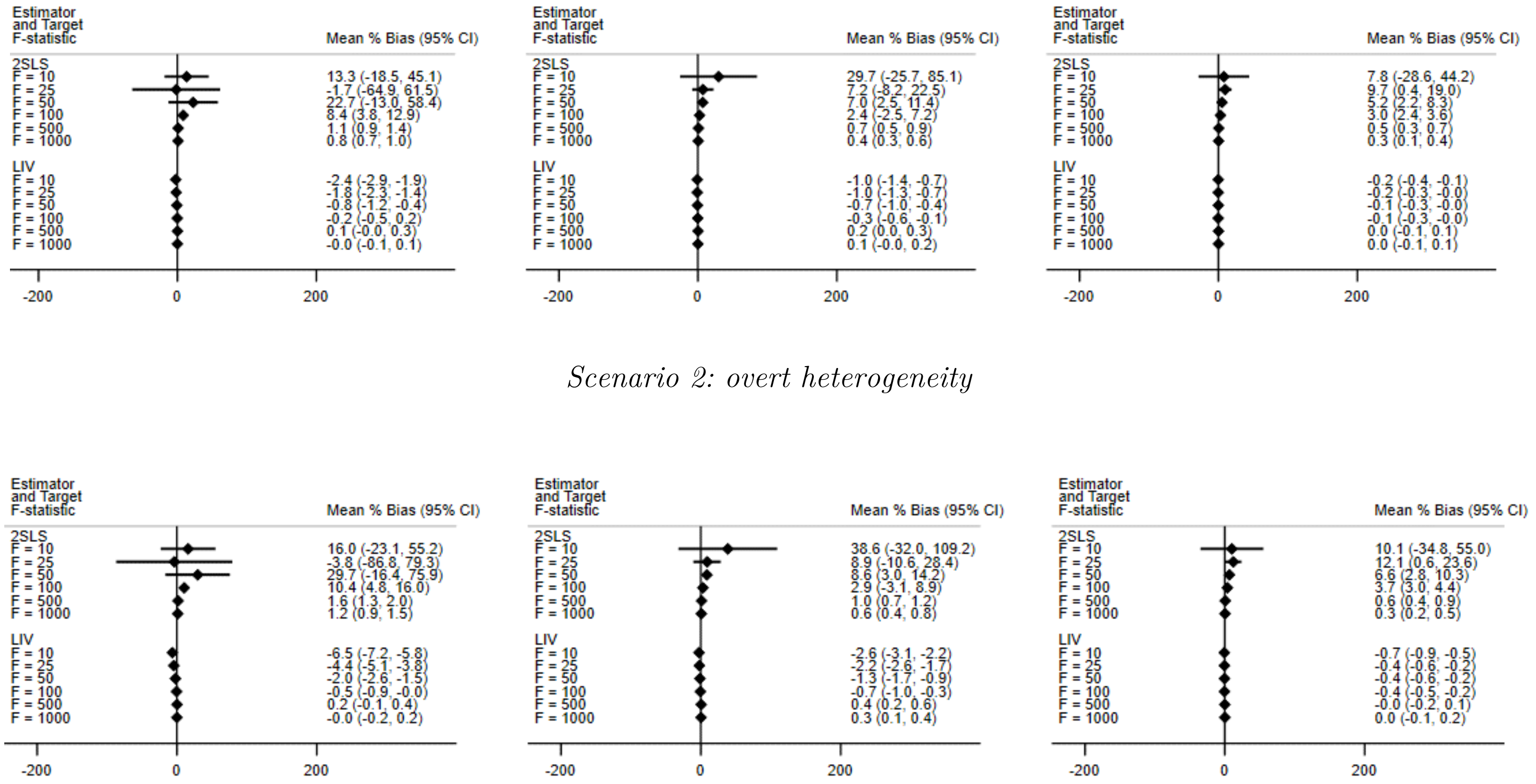
In settings with homogenous treatment effects, or with overt heterogeneity, both approaches reported relatively low levels of bias (<5%) in the ATE estimates, apart from 2SLS, which reported moderate levels of bias (5-10%) in settings with F-statistics below 100 or a smaller sample size (n = 5000) (Figure 1). In settings with essential heterogeneity, 2SLS reports relatively high (>10%) levels of mean bias across practically all combinations of IV strength and sample size. The mean (%) bias is quite variable with respect to the target F-statistic (Figure 5.1). Inspection of the distribution of percentage bias across the 5,000 simulations (not shown) suggests this is due to the fact that the tails of the distribution are fat, particularly at lower values of F. At very high (>100) levels of the target F, the mean and mean % bias are similar however this is not the case at lower levels. LIV estimator reports low levels of bias in ATE estimates across all scenarios aside from those with both a smaller sample size (n = 5000) and a F-statistic of 10 or 25 (Figure 5.1). The distribution of bias across simulation runs (not shown) has thinner tails for the LIV method than seen for 2SLS, hence the mean bias is less volatile here.

The bias plots for the CATE estimates have a somewhat similar pattern, although for this estimand the 2SLS estimator reports high levels of mean bias even in settings with overt heterogeneity, unless the sample size is relatively large (n = 50000) and/or the F-statistic is above 100 (Figure 2). The LIV estimator reports lower levels of bias than 2SLS across the majority of scenarios.

In general, for both methods, across most scenarios, for a given sample size, the levels of mean (%) bias decrease at higher levels of the F-statistic (Figure 5.2). The RMSE in the estimates of the ATE are substantially lower for the LIV than the 2SLS estimator, except for those settings with an F-statistic of 500 or 1000 (Figure 5.3). For the CATE, in general, the RMSE estimates mirror the bias results, in that they are substantially lower across all settings for LIV (Figure 5.4).

Figure 5.1. Bias plot for Average Treatment Effect (ATE) estimates across scenarios, with sample sizes of 5000 (left), 10000 (middle) and 50000 (right)

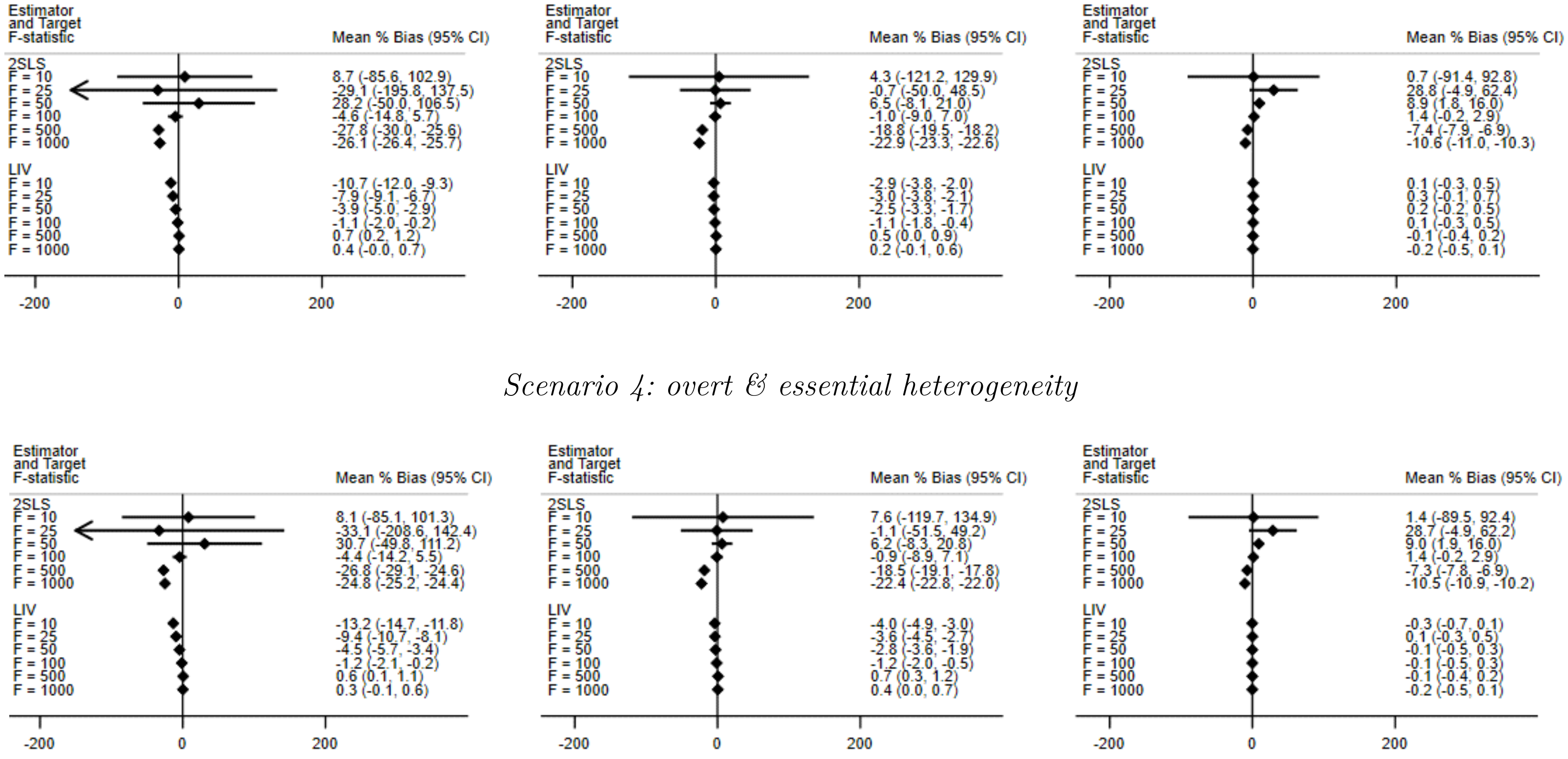
Scenario 1: effect homogeneity



Scenario 2: overt heterogeneity

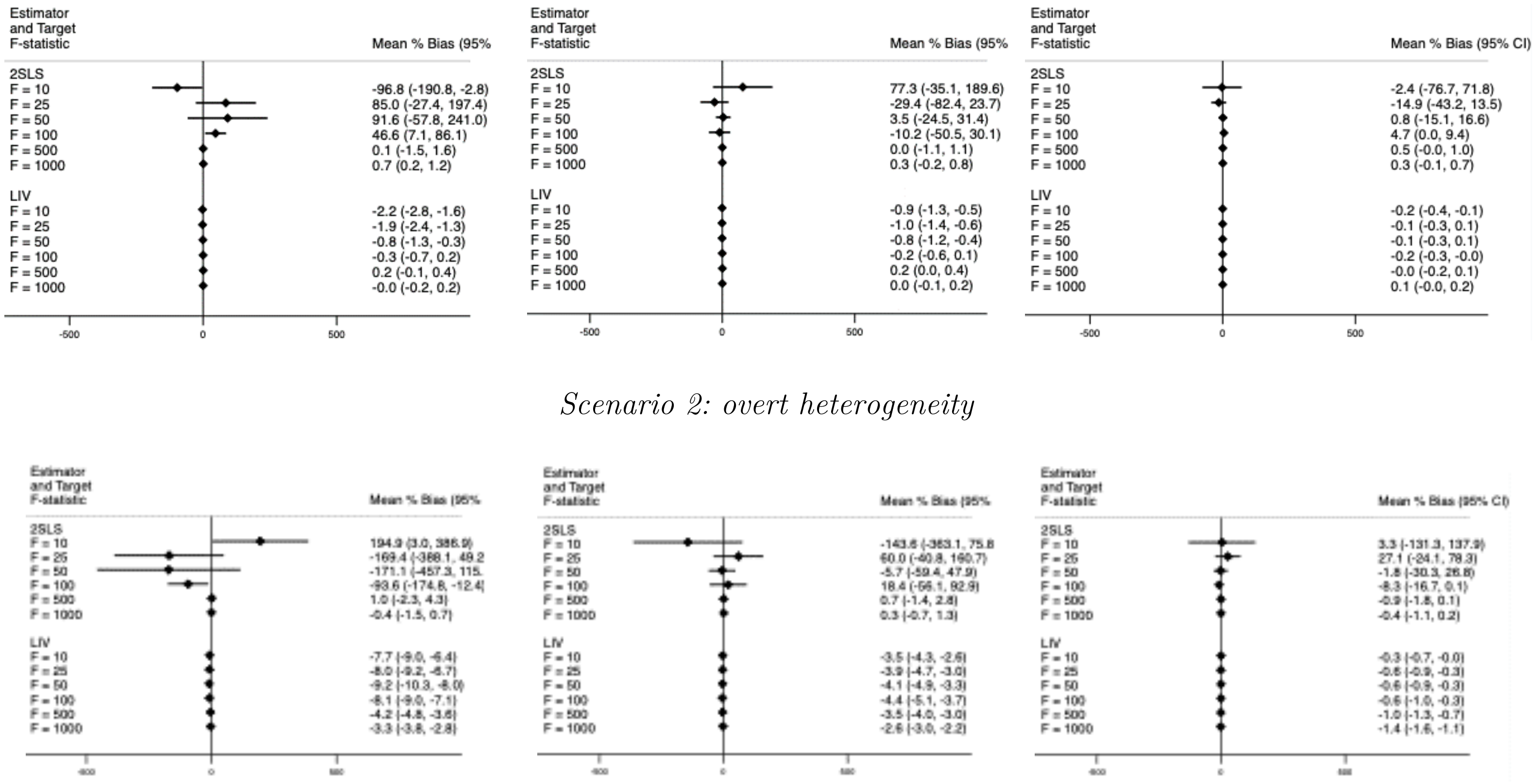
Figure 5.1. (cont.) Bias plot for Average Treatment Effect (ATE) estimates across scenarios, with sample sizes of 5000 (left), 10000 (middle) and 50000 (right)

Scenario 3: essential heterogeneity



Scenario 4: overt & essential heterogeneity

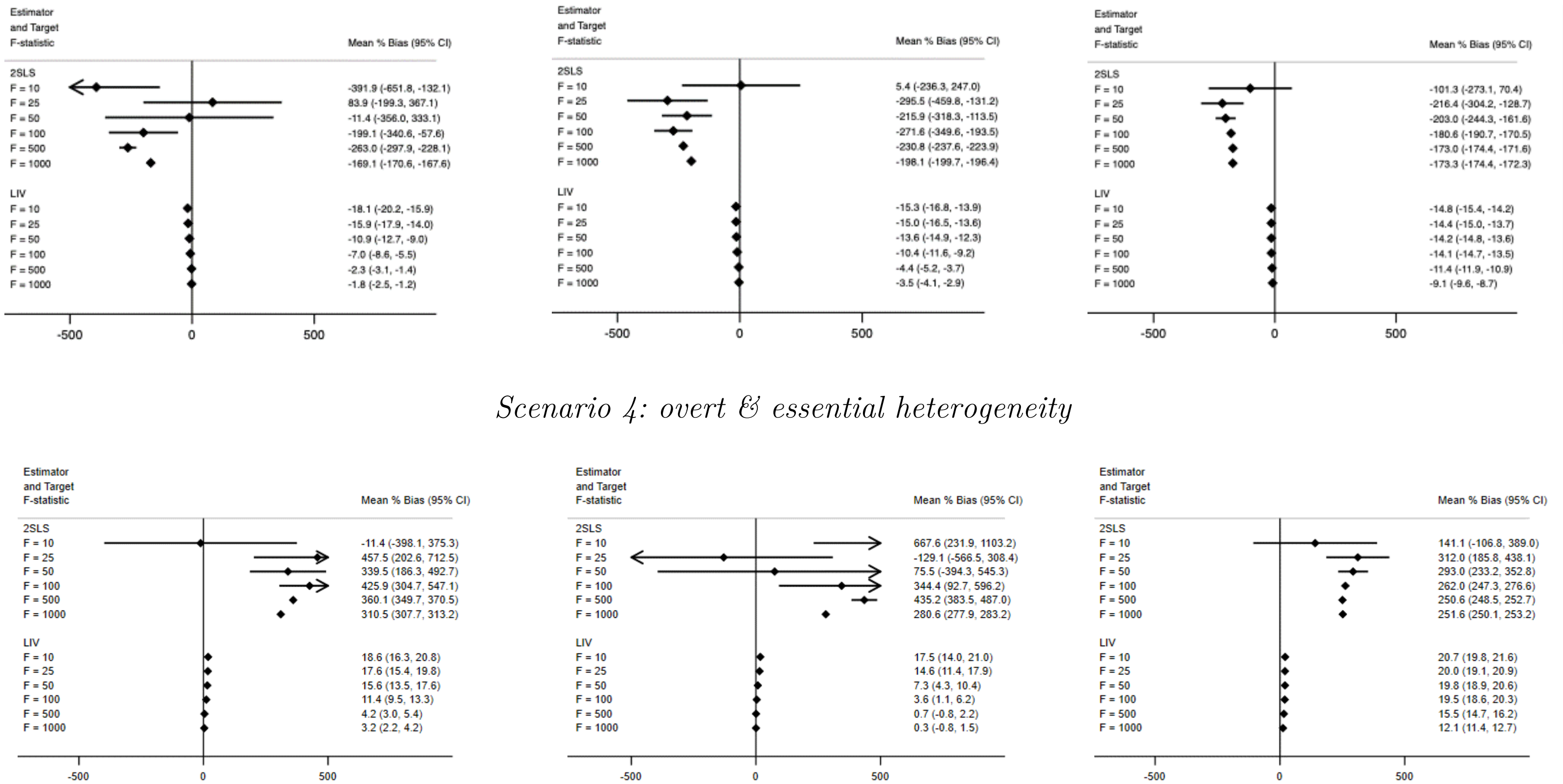
Figure 5.2. Bias plot for Conditional Average Treatment Effect (CATE) estimates across scenarios, with sample sizes of 5000 (left), 10000 (middle) and 50000 (right).



Scenario 1: effect homogeneity

Scenario 2: overt heterogeneity

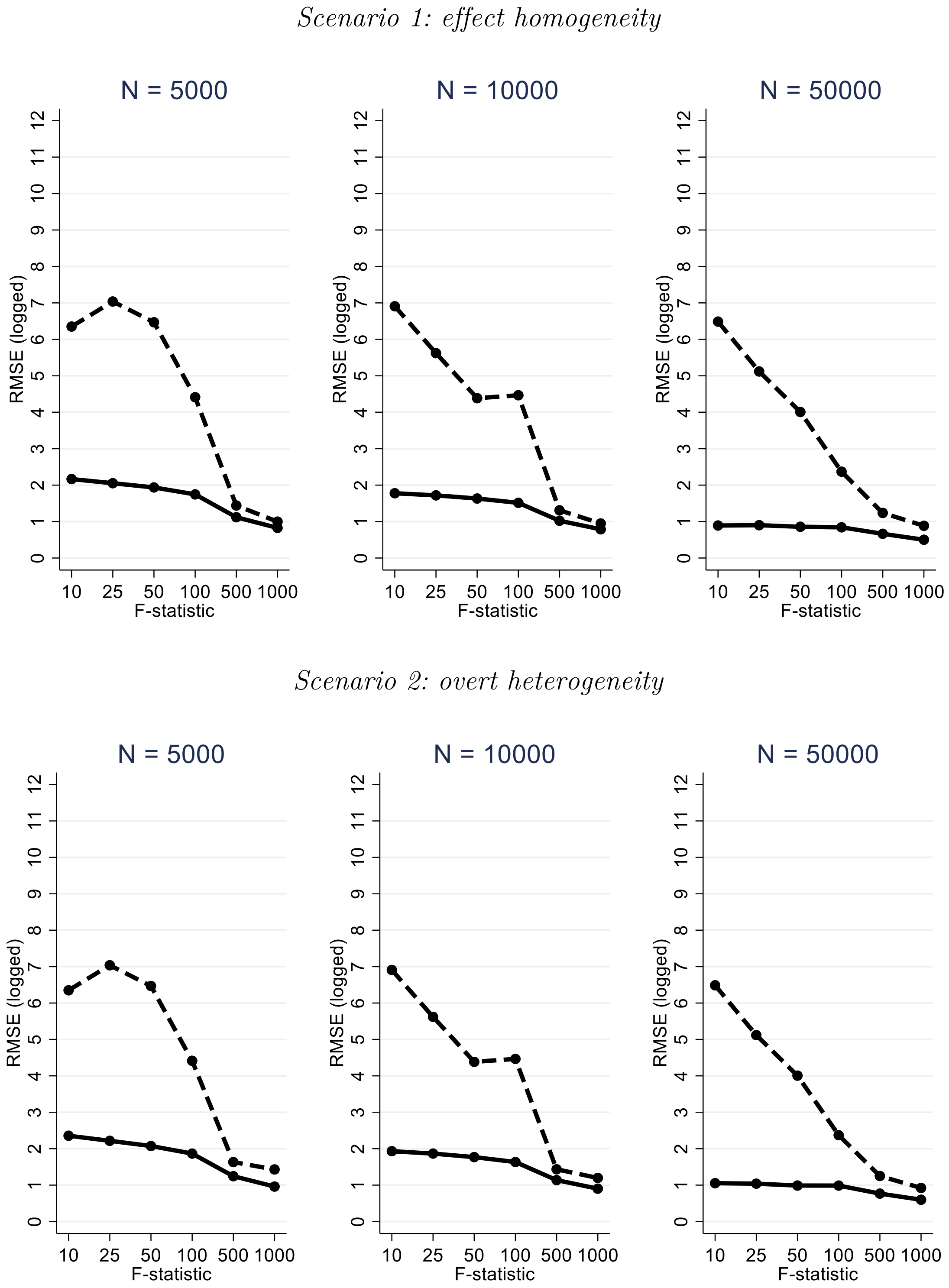
Figure 4.2. (cont.) Bias plot for Conditional Average Treatment Effect (CATE) estimates across scenarios, with sample sizes of 5000 (left), 10000 (middle) and 50000 (right)



Scenario 3: essential heterogeneity

Scenario 4: overt & essential heterogeneity

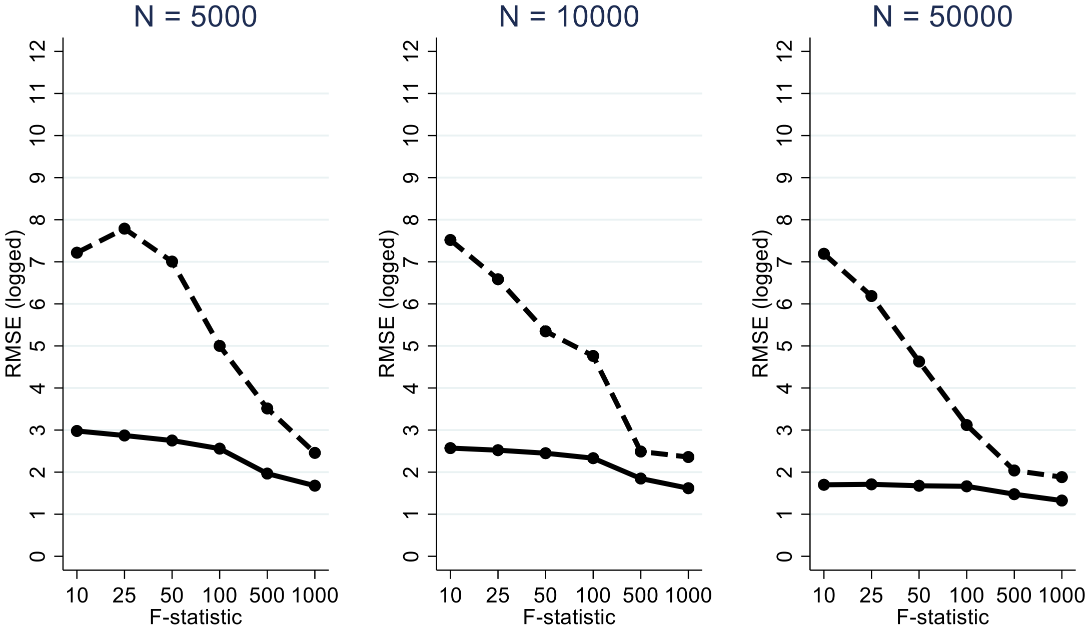
Figure 5.3. Root Mean Squared Error plots for Average Treatment Effect (ATE) estimates from 2SLS (dashed line) and LIV (solid line) across the scenarios


Scenario 1: effect homogeneity

Scenario 2: overt heterogeneity

Figure 5.3. (cont.) Bias plot for Conditional Average Treatment Effect (CATE) estimates across scenarios, with sample sizes of 5000 (left), 10000 (middle) and 50000 (right)

Scenario 3: essential heterogeneity



Scenario 4: overt & essential heterogeneity

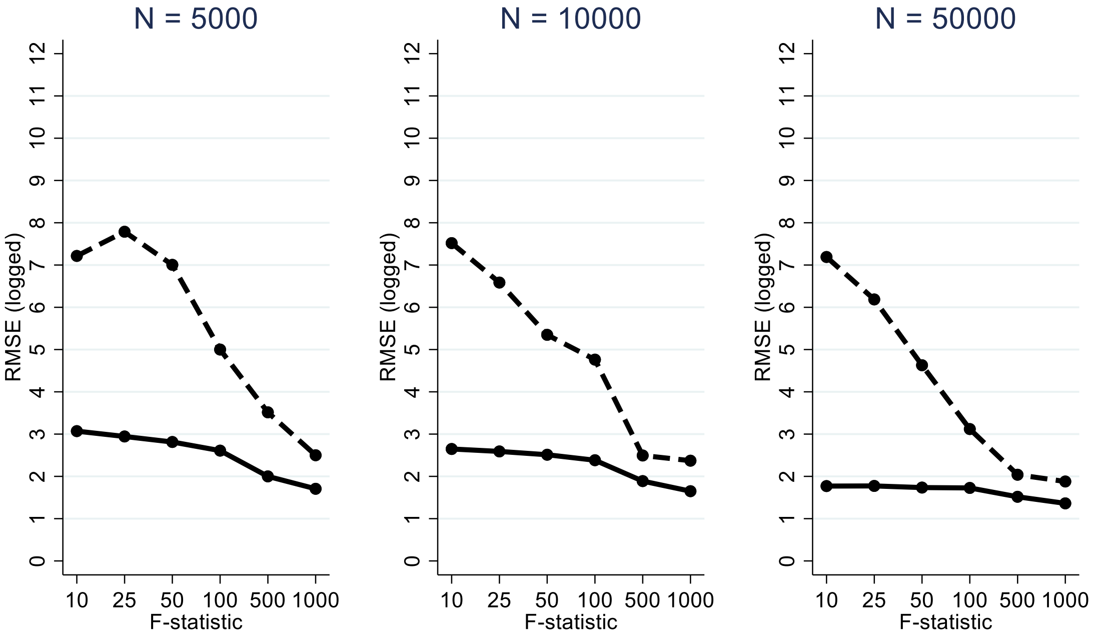
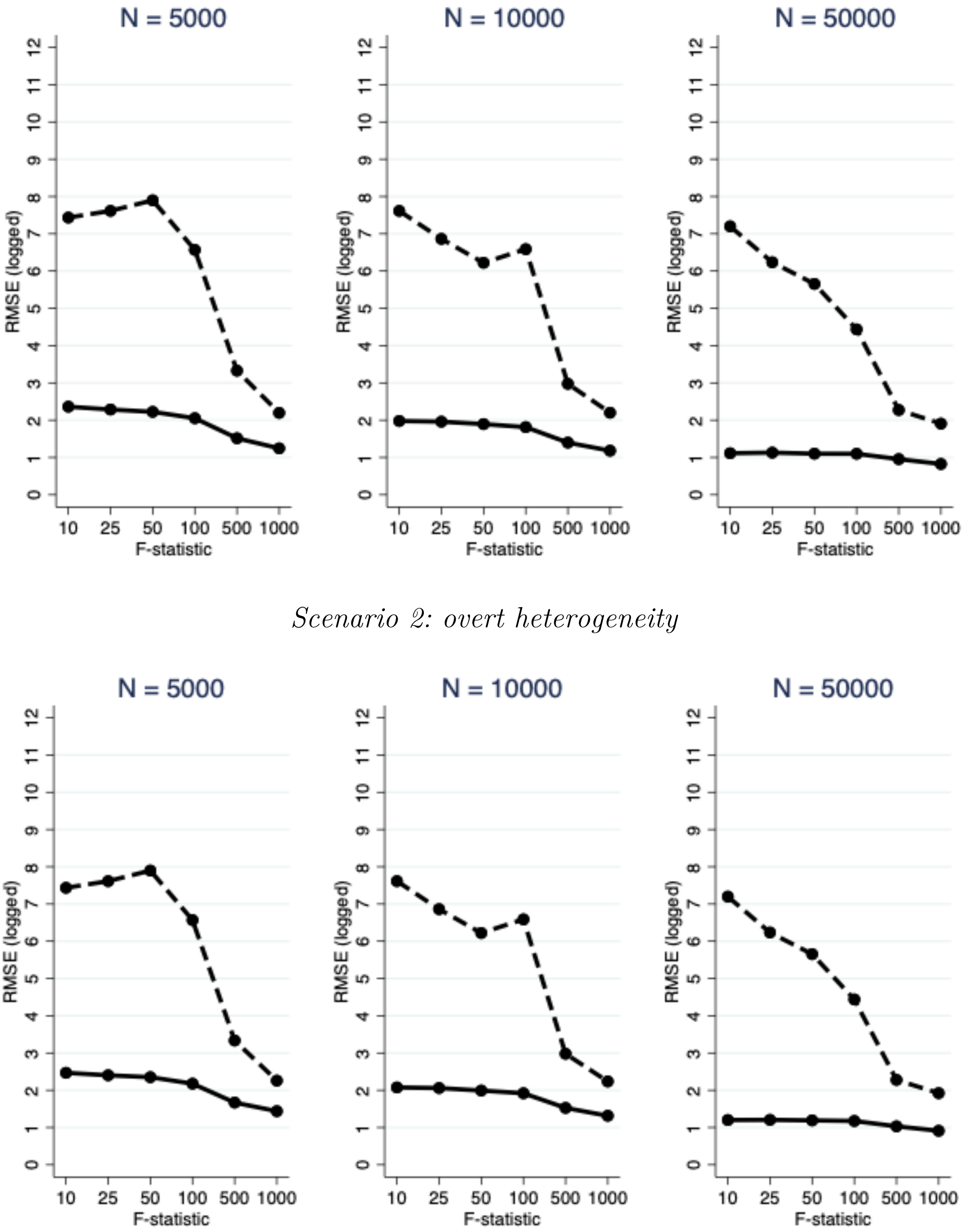


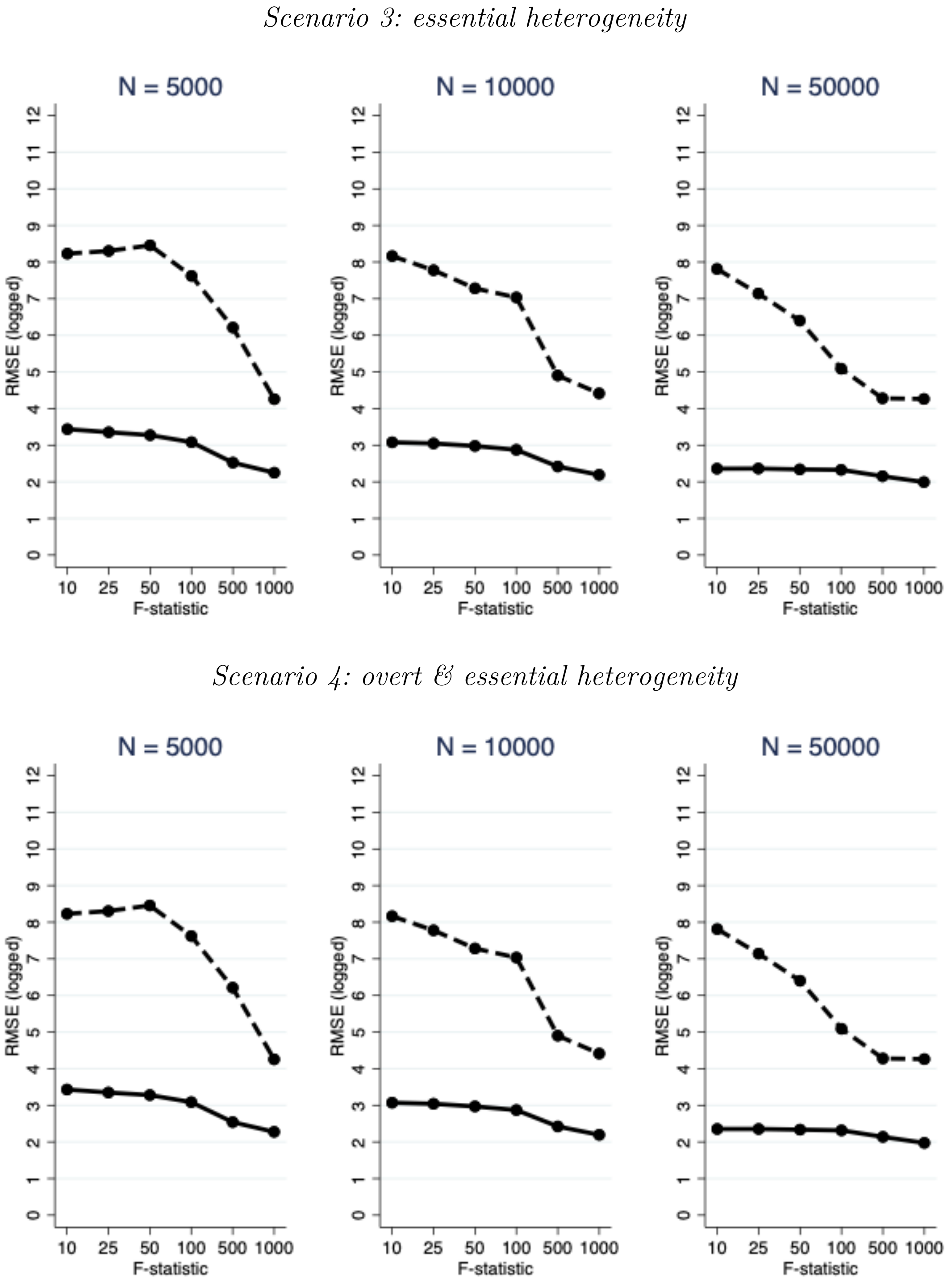
Figure 5.4. Root Mean Squared Error plots for Conditional Average Treatment Effect (CATE) estimates from 2SLS (dashed line) and LIV (solid line) across the scenarios



Scenario 1: effect homogeneity

Scenario 2: overt heterogeneity

Figure 5.4. (cont.). Root Mean Squared Error plots for Conditional Average Treatment Effect (CATE) estimates from 2SLS (dashed line) and LIV (solid line) across the scenarios



Scenario 3: essential heterogeneity

Scenario 4: overt & essential heterogeneity

Compliance rates for a given F-statistic were sensitive to the sample size available. For a sample size of 5000, increasing the F-statistic from 10 to 1000 increases the compliance rate from 8% to 73%, while for a sample size of 50000, the compliance rate only increases from 3% to 29% (Table 5.2).

Table 5.2. Compliance rate by sample size (N) and F-statistic

F-statistic	N = 5000	N = 10000	N = 50000
10	8%	6%	3%
25	13%	9%	5%
50	18%	13%	6%
100	26%	20%	9%
500	56%	42%	21%
1000	73%	57%	29%

5.2.5.2 Case study

5.2.5.2.1 Case study: implementation of 2SLS and LIV approaches

LIV estimated PeT effects of ES versus NES on DAOH at 90 days, for each individual allowing for treatment effect heterogeneity and confounding (Angrist and Kolesár, 2021; ESORT Study Group, 2020; Hutchings et al., 2022; Moffitt and Zahn, 2022). These PeT effects were aggregated to report the effects of ES overall, and for each pre-specified subgroup of interest. Since DAOH at 90 days was left skewed due to the maximum being 90 days, we rescaled this to lie between 0 and 1 (90-DAOH)/90) and effects were then rescaled back to the original scale. Probit regression models were used to estimate the initial propensity score (first stage), while GLMs were applied to the outcome data, with the most appropriate family and link function chosen according to RMSE, with Hosmer-Lemeshow and Pregibon tests also used to check model fit and appropriateness (Hosmer and Lemeshow, 2000; Pregibon, 1980). The logit link and binomial family were selected for all three conditions. Models at both stages adjusted for baseline measures, time period, and proxies for hospital quality, defined by rates of emergency readmission and mortality in 2009-10 (time constant), and in the year prior to the specific admission concerned (time-varying).

Estimates of mean differences in DAOH between the comparison groups, overall and for pre-specified subgroups (CATEs) were reported with standard errors and confidence intervals (CI) obtained with the non-parametric bootstrap (300 replications), allowing for the clustering of individuals within hospitals. The 2SLS approach used the same model specification and selection (including covariates used for confounding adjustment) to report estimates overall and for subgroups.

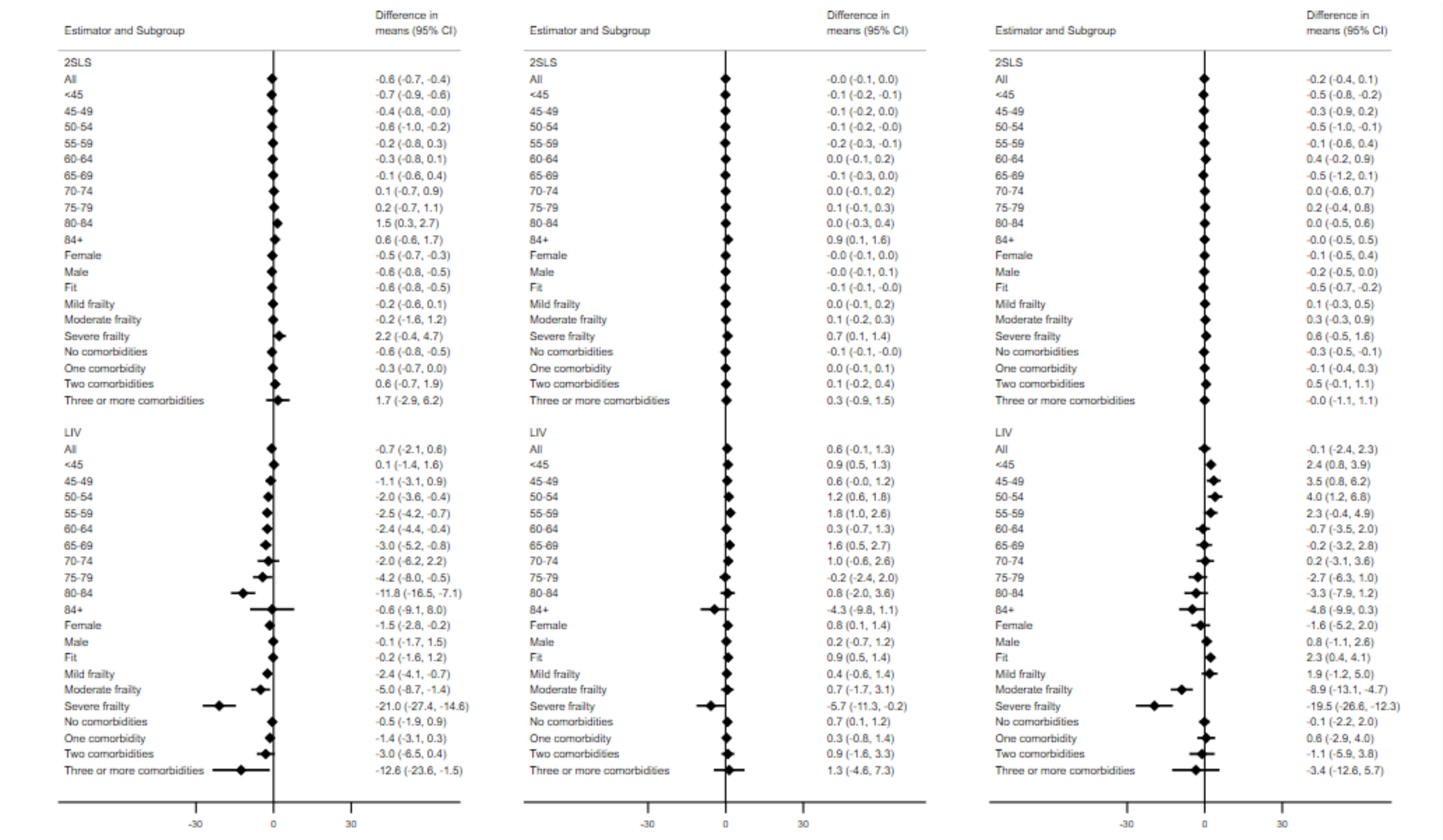
5.2.5.2.2 Case study: results

The study reported somewhat similar that for both methods the 95% CIs surrounding the mean differences included zero (Figure 5.5). Beneath this overall result, the LIV approach reported evidence that the effectiveness of ES was heterogeneous according to pre-specified subgroups. In particular, for all three conditions, ES led to lower DAOH for patients who had severe levels of frailty, and for those with acute appendicitis, ES was less effective for older patients (aged 80-84) or those with three of more comorbidities. By contrast, the 2SLS approach, which failed to account for unobserved heterogeneity (e.g., disease severity), did not report any substantive differences in relative effectiveness according to patient subgroup (Figure 5.5).

5.2.6 Discussion

This paper formally assessed the performance of the LIV methodology developed by Heckman and Vytlacil (1999, 2001) and further extended by Basu (2014) to provide policy relevant estimates of ATE and CATE in settings that differed according to the form of heterogeneity, the sample size, and level of IV strength. We contrasted the performance of LIV with that of the widely-used 2SLS approach. The scenarios considered in the simulation study were directly motivated by gaps in the literature and by a comparative effectiveness study that used LIV in evaluating emergency surgery for three acute gastrointestinal conditions for subgroups of prime policy relevance. In the case study, overt and essential heterogeneity were important concerns, amid differing levels of IV strength and sample sizes, and these issues motivated the scenario of prime interest for the simulation study (Scenario D). However, we also considered scenarios, which can, in principle provide accurate estimates of ATE and CATEs with conventional IV methods such as 2SLS (Scenarios A and B). We compared the performance of the two methods, according to bias and statistical efficiency (RMSE).

Figure 5.5. Mean differences in days alive and out of hospital (DAOH) between ES and NES for appendicitis (left), gallstone disease (centre) and hernia (right) subgroups



2SLS: two-stage least squares; CI: Confidence Interval; LIV: Local Instrumental variables.

Four preliminary findings of the simulation study are worth emphasising. First, our results suggest that while LIV performs better according to increasing levels of IV strength and sample size, this estimator reports relatively low levels of bias in estimates of the ATE and CATEs across all scenarios including those with essential heterogeneity. These findings compliment those of Basu (2014) in evaluating the reliance of the estimator on the relevance condition as well as the consistency of the estimator, but also by considering a wider range of assumptions about heterogeneity.

Second, our results suggest that 2SLS reports biased estimates of the ATE and CATEs in the presence of essential heterogeneity, except in those cases where the instrument is very strong (F-statistic above 500). These results are consistent with previous findings that 2SLS estimates cannot generally be extrapolated to broader populations beyond the compliers unless restrictive assumptions are made about the heterogeneity of treatment effects (Brooks et al., 2018; Chapman and Brooks, 2016)., However, our results suggest that, even under homogenous treatment effects, 2SLS provides biased estimates of the ATE, in scenarios where the F-statistic is low, but the requisite magnitude of the F-statistic also depends on the sample size and the form of heterogeneity.

This finding further emphasises the inadequacy of guidance resting solely on a ‘rule of thumb’ for a single setting, the target F-statistic, and highlights the importance of these wider considerations when interpreting a study’s results.

Thirdly, while 2SLS can reliably estimate CATEs in the presence of effect homogeneity or overt heterogeneity given a sufficiently strong IV or large enough sample, in the presence of essential heterogeneity, as theory would suggest, 2SLS can give extremely biased estimates of CATEs, and so in settings where essential heterogeneity is anticipated, 2SLS should not be used to estimate CATEs. In contrast, the LIV method provided estimates with low bias in the presence of overt and/or essential heterogeneity, provided the F-statistic was greater than 50. Interestingly, for the estimates of the CATEs, we find that as the sample size increases, an increase in the F-statistic is less beneficial in mitigating bias and reducing RMSE, in line with the observation that a given increase in the F-statistic has less impact on compliance rates at larger sample sizes.

Finally, LIV generally reported lower levels of RMSE than 2SLS, in particular for estimating the CATEs. However, it is important to note that here the propensity score and outcome models underlying the LIV method are correctly specified, and that performance may deteriorate where this is not the case. Data adaptive approaches could prove useful where model specification is not known.

The findings from the simulation study are informative in interpreting the CATE estimates in the ESORT study. The results offer reassurance that in such settings where essential heterogeneity would appear inevitable, that a LIV approach can provide unbiased estimate of policy-relevant estimands such as CATE, with sample sizes and F-statistics smaller than those of the ESORT study. Here, the LIV approach was able to report relative effectiveness according to subgroup, and the finding that for patients with high levels of frailty ES was not cost-effective (or cost-effective), provides important evidence to inform policy, and contributes to shared decision-making (Moler-Zapata et al., 2022).

This study has several strengths. First, it builds on insights and hypotheses raised by a large observational study using EHRs from England. The ESORT study illustrates the main challenges of using LIV methods for comparative effectiveness research and its findings in relation to IV strength, sample size requirements directly informed the scenarios considered in the simulation study. Second, while the uptake of LIV methods has been limited almost entirely to settings with essential heterogeneity, the simulation study considers different forms of heterogeneity of treatment effects as well as the scenario where treatment effects are assumed to be homogeneous in the study population. Future work will expand the simulation study to incorporate other well-known issues of IVs methods, including the challenges in applying IV estimation methods to non-linear outcome data (Clarke and Windmeijer, 2010; Vansteelandt et al., 2011). Previous research has shown that the power of 2SLS conveyed by conventional F-statistic values is low (Keane and Neal, 2021; Lee et al., 2021). In this future work, we will therefore consider the implications of sample size and instrument strength for the power of LIV analyses and confidence interval coverage. Future work will also formally assess whether imbalances in treatment assignment rates are detrimental to consistency and power of LIV inferences. This is an important concern for applied work using EHRs. For instance, the observed difference in the prevalence of ES and NES in ESORT (90/10 in the cohort with appendicitis) could reduce the power of the analysis (Walker et al., 2017).

Acknowledgements

We would like to acknowledge Professor Luke Keele for helpful discussions, and other members of the ESORT study team. We would like to acknowledge the feedback from those who participated in the session and particularly the paper discussant at the HESG Summer 2022 conference, Dr Manuel Gomes.

References

Abadie A (2003) *Semiparametric Instrumental Variable Estimation of Treatment Response Models*. DOI: 10.1016/S0304-4076(02)00201-4.

Andrews I, Stock JH and Sun L (2019) Weak Instruments in Instrumental Variables Regression: Theory and Practice. *Annual Review of Economics* 11: 727–753. DOI: 10.1146/annurev-economics-080218-025643.

Angrist J and Kolesár M (2021) One Instrument to Rule Them All: The Bias and Coverage of Just-Id IV. *SSRN Electronic Journal*. DOI: 10.2139/ssrn.3953944.

Angrist J, Imbens G and Rubin D (1993) Identification of causal effects using instrumental variables. *Journal of the American Statistical Association* 91(434): 444–455.

Angrist JD and Fernández-Val I (2011) ExtrapoLATE-ing: External validity and overidentification in the LATE framework. *Advances in Economics and Econometrics: Tenth World Congress Volume 3, Econometrics*: 401–434. DOI: 10.1017/CBO9781139060035.012.

Baiocchi M, Cheng J and Small DS (2014) Instrumental variable methods for causal inference. *Statistics in Medicine* 33(13): 2297–2340. DOI: 10.1002/sim.6128.

Basu A (2014) Estimating person-centered treatment (PeT) effects using instrumental variables: an application to evaluating prostate cancer treatments. *JOURNAL OF APPLIED ECONOMETRICS* 29: 671–691. DOI: 10.1002/jae.

Basu A (2015) Person-centered treatment (PeT) effects: Individualized treatment effects using instrumental variables. *The Stata Journal* 15(2): 397–410.

Basu A and Rathouz PJ (2005) Estimating marginal and incremental effects on health outcomes using flexible link and variance function models. *Biostatistics* 6(1): 93–109. DOI: 10.1093/biostatistics/kxh020.

Basu A, Heckman JJ, Navarro-Lozano S, et al. (2007) Use of instrumental variables in the presence of heterogeneity and self-selection: an application to treatments of breast cancer patients. *Health Economics* 16(2007): 1133–1157. DOI: 10.1002/hec.1291.

Basu A, Coe NB and Chapman CG (2018) 2SLS versus 2SRI: Appropriate methods for rare outcomes and/or rare exposures. *Health Economics* 27(6): 937–955. DOI: 10.1002/hec.3647.

Basu A, Jones AM and Rosa Dias P (2018) Heterogeneity in the impact of type of schooling on adult health and lifestyle. *Journal of Health Economics* 57. Elsevier B.V.: 1–14. DOI: 10.1016/j.jhealeco.2017.10.007.

Bjorklund A and Moffitt R (1983) *Estimation of Wage Gains and Welfare Gains from Self-Selection Models. IUI Working Paper, No. 105,*. Stockholm.

Bound J, Jaeger DA and Baker RM (1995) Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak. *Journal of the American Statistical Association* 90(430): 443–450. DOI: 10.1080/01621459.1995.10476536.

Brooks JM, Chapman CG and Schroeder MC (2018) Understanding Treatment Effect Estimates When Treatment Effects Are Heterogeneous for More Than One Outcome. *Applied Health Economics and Health Policy* 16(3): 381–393. DOI: 10.1007/s40258-018-0380-z.

Chapman CG and Brooks JM (2016) Treatment Effect Estimation Using Nonlinear Two-Stage Instrumental Variable Estimators: Another Cautionary Note. *Health Services Research* 51(6). Blackwell Publishing Inc.: 2375–2394. DOI: 10.1111/1475-6773.12463.

Clarke P and Windmeijer F (2010) Instrumental Variable Estimators for Binary Outcomes. *CMPO Working Paper Series No. 10/239 Instrumental*. Available at: http://www.bristol.ac.uk/cmpo/Tel:

Cornelissen T, Dustmann C, Raute A, et al. (2018) Who benefits from universal child care? Estimating marginal returns to early child care attendance. *The Journal of political economy* 126(6): 2356–2409. Available at: http://www.christiandustmann.com/content/4-research/2-who-benefits-from-universal-childcare-estimating-marginal-returns-to-early-childcare-attendance/cornelissen_etal_2017_jpe_forthcoming.pdf.

ESORT Study Group (2020) Emergency Surgery Or NoT (ESORT) study. Available at: https://www.lshtm.ac.uk/media/38711.

Grieve R, O'Neill S, Basu A, et al. (2019) Analysis of Benefit of Intensive Care Unit Transfer for Deteriorating Ward Patients: A Patient-Centered Approach to Clinical Evaluation. *JAMA network Open* 2(2). NLM (Medline): 1–13. DOI: 10.1001/jamanetworkopen.2018.7704.

Heckman JJ and Vytlacil E (2005) Structural equations, treatment effects and econometric policy evaluation. *Econometrica* 73(3): 669–738. DOI: 10.1111/j.1468-0262.2005.00594.x.

Heckman JJ and Vytlacil EJ (1999) Local instrumental variables and latent variable models for identifying and bounding treatment effects. *Proceedings of the National Academy of Sciences of the United States of America* 96: 4730–4734. DOI: 10.1073/pnas.96.8.4730.

Heckman JJ and Vytlacil EJ (2001) Policy-Relevant Treatment Effects. *American Economic Review* 91(2): 107–111. DOI: 10.1257/aer.91.2.107.

Heckman JJ, Urzua S and Vytlacil E (2006) *Understanding instrumental variables in models with essential heterogeneity. NBER Working Paper No. 12574.* Cambridge, MA. DOI: 10.1162/rest.88.3.389.

Hosmer DW and Lemeshow S (2000) *Applied Logistic Regression.* 2nd ed. Wiley.

Hutchings A, Moler-Zapata S, O'Neill S, et al. (2021) Variation in the rates of emergency surgery amongst emergency admissions to hospital for common acute conditions. *BJS Open*. DOI: 10.1093/bjsopen/zrab094.

Hutchings A, O'Neill S, Lugo-palacios DG, et al. (2022) Effectiveness of emergency surgery for five common acute conditions: an instrumental variable analysis of a national routine database. *Anaesthesia*: In Press.

Imbens GW and Angrist JD (1994) Identification and Estimation of Local Average Treatment Effects. *Econometrica* 62(2). JSTOR: 467. DOI: 10.2307/2951620.

Keane M and Neal T (2021) *A Practical Guide to Weak Instruments. UNSW Economics Working Paper No. 2021-05d.*

Keele L, Sharoky CE, Sellers MM, et al. (2018) An instrumental variables design for the effect of emergency general surgery. *Epidemiologic Methods* 7(1). Walter de Gruyter GmbH. DOI: 10.1515/em-2017-0012.

Lee D, McCrary J, Moreira MJ, et al. (2021) *Valid T-Ratio Inference for IV. National Bureau of Economic Research Working Paper Series (No. w29124).* DOI: 10.2139/ssrn.3901588.

Little RJ, Long Q and Lin X (2009) A comparison of methods for estimating the causal effect of a treatment in randomized clinical trials subject to noncompliance. *Biometrics* 65(2): 640–649. DOI: 10.1111/j.1541-0420.2008.01066.x.

Martínez-Camblor P, MacKenzie TA, Staiger DO, et al. (2019) An instrumental variable procedure for estimating Cox models with non-proportional hazards in the presence of unmeasured confounding. *Journal of the Royal Statistical Society. Series C: Applied Statistics* 68(4): 985–1005. DOI: 10.1111/rssc.12341.

Moffitt RA and Zahn M V (2022) The Marginal Labor Supply Disincentives of Welfare : Evidence from Administrative Barriers to Participation.

Moler-Zapata S, Grieve R, Lugo-Palacios D, et al. (2022) Local instrumental variable methods to address confounding and heterogeneity when using electronic health records: an application to emergency surgery. *Medical Decision Making* 0(0). DOI: 10.1177/0272989X221100799.

Nelson CR and Startz R (1990) Some further results on the exact small sample properties of the instrumental variables estimator. *Econometrica*: 967–976.

Neyman J (1990) On the application of probability theory to agricultural experiments. *Statistical Science* 5: 463–480.

Pregibon D (1980) Goodness of Link Tests for Generalized Linear Models. *Journal of the Royal Statistical Society. Series C (Applied Statistics),* 29(1). London: Royal Statistical Society: 14–15. DOI: 10.2307/2346405.

Reynolds K, Barton LJ, Basu A, et al. (2021) Comparative Effectiveness of Gastric Bypass and Vertical Sleeve Gastrectomy for Hypertension Remission and Relapse: The ENGAGE CVD Study. *Hypertension* 78(4): 1116–1125. DOI: 10.1161/HYPERTENSIONAHA.120.16934.

Rubin DB (1974) Estimating causal effects of treatment in randomized and nonrandomized studies. *Journal of Educational Psychology* 66(5): 688–701. Available at: http://www.fsb.muohio.edu/lij14/420_paper_Rubin74.pdf.

Small DS and Rosenbaum PR (2008) *War and Wages: The Strength of Instrumental Variables and Their Sensitivity to Unobserved Biases.* DOI: 10.1198/016214507000001247.

Staiger D and Stock JH (1997) Instrumental Variables Regression with Weak Instruments. *Econometrica* 65(3): 557. DOI: 10.2307/2171753.

Stata Corp Lp (2001) Stata Statistical Software: Release 7.0. College Station, TX. Stata Press Publication.

Stock JH, Wright J and Yogo M (2002) GMM , Weak Instruments , and Weak Identification. *Journal of Business and Economic Statistics symposium.*

Tan Z (2006) Regression and weighting methods for causal inference using instrumental variables. *Journal of the American Statistical Association* 101(476): 1607–1618. DOI: 10.1198/016214505000001366.

Terza J V., Bradford WD and Dismuke CE (2008) The use of linear instrumental variables methods in health services research and health economics: A cautionary note. *Health Services Research* 43(3). John Wiley & Sons, Ltd: 1102–1120. DOI: 10.1111/j.1475-6773.2007.00807.x.

Terza J V., Basu A and Rathouz PJ (2008) Two-stage residual inclusion estimation: Addressing endogeneity in health econometric modeling. *Journal of Health Economics* 27(3). NIH Public Access: 531–543. DOI: 10.1016/j.jhealeco.2007.09.009.

Vansteelandt S, Bowden J, Babanezhad M, et al. (2011) On instrumental variables estimation of causal odds ratios. *Statistical Science* 26(3): 403–422. DOI: 10.1214/11-STS360.

Vytlacil E (2002) Independence, monotonicity and latent index models: an equivalence result. Econometrica 70(1): 331–341. DOI: 10.1080/0305006790150101.

Walker VM, Davies NM, Windmeijer F, et al. (2017) Power calculator for instrumental variable analysis in pharmacoepidemiology. *International Journal of Epidemiology* 46(5): 1627–1632. DOI: 10.1093/ije/dyx090.

Chapter 6. Discussion

6.1 Introduction

Comparative effectiveness and cost-effectiveness studies that evaluate alternative health and social care interventions have the opportunity to exploit the growing availability of RWD, and provide evidence that can inform decision-making. However, these studies are faced with major challenges including the risk of unmeasured confounding and heterogeneity. A further concern and barrier to the wider use of this form of evidence in decision-making, is the lack of transparency about choices in the study design and analysis. These concerns may apply to any setting that uses RWD to assess comparative effectiveness and cost-effectiveness, including pharmaceutical interventions, other health technologies such as surgical interventions or devices, changes to health or care services or the introduction of new health policies, or public health interventions (Faria et al., 2015; Skivington et al., 2021). This thesis draws on recent methods developments in the causal inference and health econometrics literature to help improve the approaches for tackling confounding and heterogeneity when assessing comparative effectiveness and cost-effectiveness using RWD.

The aim of this thesis was to help address gaps in the guidance on methods for CEA that use RWD, and more specifically routine data, in settings with unmeasured confounding and treatment effect heterogeneity. The specific objectives were to:

1. Critically examine the application of the principles of the target trial framework to the HTA context, identify the main challenges, and provide recommendations to address them.
2. Evaluate and implement an LIV approach for addressing unmeasured confounding and heterogeneity in CEA.
3. Evaluate the performance of IV approaches in terms of bias and statistical efficiency according to alternative levels of IV strength, sample sizes and forms of heterogeneity in a simulation study.

The next section outlines the main findings from the thesis. Sections 6.3 and 6.4 discuss the main contributions of the thesis to the methodological literature, and the literature evaluating the cost-effectiveness of ES for acute gastrointestinal conditions. Section 6.5 summarises the main limitations of the thesis. Section 6.6 identifies areas for future research. Section 6.7 and 6.8 discuss the implications for applied researchers and policy making, and section 6.9 concludes the thesis.

6.2 Overall findings

The target trial framework can be adopted in CEA to support the design of studies that use RWD. Since the publication of the seminal papers by Hernán and Robins (2016) and Hernán et al., (2016), the target trial framework has become a popular tool for non-randomised health care evaluations as it can help reduce bias due to confounding which is a perennial problem with this study design. The target trial framework also provides evidence users, including decision-makers and service providers, with a tool to judge methodological choices of observational studies, such as the plausibility of assumptions made in the statistical analysis, according to how closely they emulate the design elements of the hypothetical trial.

This thesis identified four main challenges in applying the target trial framework within CEA that use routine data. These relate to potential data constrains affecting the study's ability to emulate the trial's eligibility criteria, challenges in defining treatment strategies and time zero, and the risk of confounding. I argue that these four challenges are prominent in those common settings in which the analyst has no control over the data collection process. These four challenges are important ones to address to help reduce bias from confounding and improve the transparency and reproducibility of methods and findings, to improve the use of RWD for informing decision-making. While these issues are not the only challenges that studies using RWD might face, other concerns may relate for example to missing data, censoring, non-compliance or measurement error, but these are beyond the scope of the thesis (DiazOrdaz and Grieve, 2019; Latimer et al., 2014; Willan and Briggs, 2006). Here, the focus of the thesis is on approaches to address the unmeasured confounding and essential heterogeneity that arise in estimating comparative effectiveness and cost-effectiveness from routine data.

Research paper 1 offers an exemplar application of the target trial framework in a CEA, and critically considers the major challenges that could arise in comparative effectiveness and cost-effectiveness studies that use RWD. The first challenge is to define the eligibility criteria that delineate the target population for the study. These are patients who would be eligible to receive either the intervention or the comparator in routine practice. However, the emulation of a trial's design should also ensure that the population only includes those patients for whom there is equipoise about the treatment comparators of interest (i.e., patients with a positive probability of receiving each treatment option). Otherwise, if the study includes a subgroup for whom there is a strong prior belief that they will not benefit from either of the comparators of interest, even after adjusting for confounding factors, the treatment effect could be estimated with bias (Petersen et al., 2010). Researchers should therefore consider the "equipoise" principle when defining the eligibility criteria. This is can be challenging as patient's equipoise often cannot be assessed from the data. The paper offers some recommendations for how to apply this principle to the available RWD, recognising that in practice there may be gaps, omissions and challenges in applying the framework in practice.

One related challenge is in defining treatment strategies from the RWD. If the definition of the intervention differs from the intervention of interest, including the setting in which they are administered or their timing, the findings of the study will be of limited relevance for decision-making. Researchers also need to carefully consider the definition of the comparator strategy. Ideally, it should reflect the comparator used in routine clinical practice to be relevant for decision-making. It is also important that the comparators are be defined in sufficient level of detail to be able the causal contrast of interest is identified (Hernán, 2004; Hernán and Taubman, 2008).

Third, the study also exemplified the concern that, even if treatment strategies and target population can be defined from the RWD, it is important to define time (day) zero the analogue to the time of randomisation, as part of any strategy to reduce bias in the estimation of treatment effects. The date or time of key events is not always recorded in the RWD, which makes it challenging for studies to define when eligibility is met, treatment is assigned and the treatment strategies are initiated (Patorno et al., 2020). The paper discusses how CEA that fail to align these events could suffer from selection bias and immortal time bias.

Lastly, the paper discusses the role that insufficient information for covariate adjustment plays in raising concerns of residual confounding in CEA. This is a well-known challenge for observational studies, and previous studies have considered alternative statistical methods for tackling confound in CEA (Kreif et al., 2013; Nixon and Thompson, 2005; Polsky and Basu, 2012; Sekhon and Grieve, 2012). However, one important (new) finding of this thesis is that IV designs can be compatible with the application of target trial framework in observational studies. Previous studies had suggested that valid IVs are inadequate for emulating trials, mainly because the estimated effect only pertains to a subset of the population (Swanson, 2017), which is very unlikely to represent the target population. While this is true in conventional IV analyses, as my second research paper showed, LIV methods can, provided some requisite assumptions hold, provide policy-relevant estimands, for example the ATE for the whole population of interest, and CATE for subpopulations of interest.

Research paper 2 directly addresses the concern that few CEA have applied IV methods to RWD to formally model essential heterogeneity in patient outcomes, costs and cost-effectiveness of an intervention. Here, the concern is that unmeasured characteristics that predict expected outcomes following either intervention also inform treatment selection. The paper contrasts different IV approaches (2SLS, 2SRI and LIV) in evaluating the cost-effectiveness of the alternative strategies, in this case ES versus alternatives. These methods have different target estimands, and different assumptions for identification, which makes comparison between the findings challenging. However, it is still interesting that the findings from the three methods are notably different. 2SLS, which aims to estimate the LATE, and LIV which aims to estimate the ATE give different estimates of the overall effect, especially in the diverticular disease cohort. The reason why estimates are so dissimilar in that cohort, might be explained by essential heterogeneity, as surgical teams may be likely to select patients into treatments according to expected gains from treatment given their unobserved risk profile. Theory suggests that both 2SRI and LIV can report estimates of the ATE, so differences in the resultant estimates are likely to reflect that in the presence of essential heterogeneity suspected in the diverticulitis example, the underlying assumptions are less plausible for the 2SRI versus LIV approach (Basu et al., 2018).

In research paper 2, I exemplify the LIV approach to estimate the overall cost-effectiveness of ES for the study population, as well as for a series of pre-specified subgroups. The study evaluated ES using different approaches: risk-adjustment (GLM regression), 2SLS, 2SRI and LIV. While all four study designs rely on untestable assumptions, only the LIV design can identify causal effects in the presence of unmeasured confounding and treatment effect heterogeneity according to unobserved characteristics. LIV approaches do still rely on the fundamental IV assumptions, and the presence of a continuous IV (see Chapter 2) (Basu et al., 2007). The plausibility of these assumptions was carefully evaluated using both formal tests and clinical judgement. The TTO was strongly associated with receipt of ES. Even though the exclusion restriction can never be fully tested, the study found that the TTO was able to balance the baseline covariates, which makes it more plausible that it was also able to balance unobserved covariates. Models were adjusted for measures of hospital quality of care to further bolster the plausibility of the identification assumption.

One important finding of the analysis using LIV is that while the study does not provide strong evidence that either ES or NES is cost-effective overall for any of the acute conditions, one or other of the modalities can be cost-effective when targeted at specific population subgroups. In particular, for patients with acute appendicitis and abdominal wall hernia with moderate or severe less of frailty, and those who have at least two comorbidities, the NES strategy is relatively cost-effective. NES is also cost-effective for diverticular disease patients with perforation of abscess. ES appears to be cost-effective for subgroups of patients with diverticular disease or hernia who are fit (both conditions), or who are younger (hernia only).

Research paper 2 also reported differences across the acute conditions in the cohort sizes and the strength of the IV (see Table 4.3). These findings motivated further research questions about the requirements of IV methods with respect to instrument strength and sample size, for estimation and inference pertaining to ATE and CATE. and helped define the scenarios of interest for the simulation study in research paper 3.

The main findings from that paper 3 are that, first LIV methods perform well regardless of the form of heterogeneity and confounding; second the general requirements with respect to IV strength depend on the available sample size; third, 2SLS estimates are biased in settings with essential heterogeneity when the instrument is not strong, fourth, levels of strength used to define sufficiency of IV strength in applications might fail to guarantee minimal biases in the estimation of treatment effects.

6.3 Contributions

This thesis contributes to the literature on analytical methods for CEA by drawing on insights from the causal inference and health econometrics literature. The following sections describe the main methodological contributions of the thesis.

6.3.1 Developing recommendations for studies that apply the target trial framework to Real-World Data

Research paper 1 offers a series of recommendations for CEA to address the main challenges in applying the target trial framework to studies that use routine patient-level data (see previous section). These recommendations complement previous methodological guidance (Drummond et al., 2005; Faria et al., 2015; Husereau et al., 2013; Philips et al., 2004) and checklists for economic evaluations, and expand on them by considering aspects of the study design that are specific to studies that use RWD. The contributions also complement previous tools developed for evaluating statistical approaches used in observational CEA, including those that were designed to address selection bias (Kreif et al., 2013), by considering broader aspects of the study design of observational studies.

Research paper 1 discusses how expert opinion can help in adapting the target trial's eligibility criteria, and definitions of the treatment strategies to the RWD available to minimise the risk of bias. One example of how expert judgement can be used, in defining which of the eligibility criteria available from within the RWD are required to ensure equipoise between treatment strategies. Previous recommendations for aligning eligibility criteria in the study with those in the target trial include measuring the proportion of patients included/excluded as a result of applying each criterion (Franklin et al., 2020). Lodi et al., (2019) describes methods to 'harmonise' the target trial design with published RCT, and to accompany these attempts with sensitivity analyses to explore the impact of components that cannot be harmonised.

In considering how the target trial framework can be applied to IV designs, research paper 1 also provides practical advice for future CEA conducted in those settings where there is a risk of unmeasured confounding. These recommendations highlight the importance of contrasting methods that make alternative assumptions about confounding, as well as carefully evaluating the presence of heterogeneous treatment effects. In these settings, even if the IV is judged valid, the requisite assumptions of IV methods like 2SLS will not be satisfied, and the resultant inferences are unlikely to be appropriate (see also next section).

6.3.2 Application of a Local Instrumental Variable approach for unmeasured confounding and heterogeneity in Cost-Effectiveness Analysis that use Real-World Data

The main contribution of research paper 2 is to the literature on CEA methods, in illustrating how LIV methods can estimate treatment effects in the presence of unmeasured confounding and treatment effect heterogeneity. It describes the assumptions required for identification of relevant treatment effect parameters, and exemplifies the issues that arise when undertaking a policy-relevant CEA that relies on routine data for estimating comparative effectiveness. The paper contrasts LIV with alternative IV approaches (2SRI and 2SLS) which make alternative assumptions, and offers guidance for future CEA on how to interpret those discrepancies.

Another contribution of research paper 2 is in demonstrating how, LIV can evaluate treatment effect heterogeneity over pre-specified subgroups of interest. This is an important contribution to the literature on methods for informing personalisation of treatment choice in clinical practice. This literature has expanded rapidly in recent years, but most methods assume that treatment selection is only according to observed covariates (Kreif et al., 2020; Sadique et al., 2022) which in many CEA that use RWD is an unrealistic assumption. For example, decision-making as to which intervention patients receive may reflect 'capacity to gain' which is likely to reflect biological, patient or physician preferences or organisational characteristics which are unlikely to be measured within the data. One extension that could be considered in future studies that use the LIV framework is to harness the myriad of covariates available from linked datasets with advances in the machine learning literature, to select those observed covariates that modify relative effectiveness and cost-effectiveness (see also section 6.6) (Belloni et al., 2014)

The findings of research paper 2 on the cost-effectiveness of ES for treating patients with acute appendicitis, diverticular disease and abdominal wall hernia (described in the previous section) also contribute to the limited available evidence on the cost-effectiveness of ES for acute conditions from previous RCTs and observational studies. Compared to previous studies for the three conditions, the CEA considers broader more heterogeneous populations (O'Leary et al., 2021) evaluates economic outcomes, including resource use and costs (Flum et al., 2020) and reports heterogeneity in effects, costs and cost-effectiveness of ES versus NES strategies for each of these acute conditions (Javanmard-Emamghissi et al., 2021).

6.3.3 Evaluation of instrument strength requirements for Local Instrumental Variables in simulation study

This thesis has demonstrated that there is great potential in applying LIV in CEA and comparative effectiveness studies more generally. One important contribution is to the health econometrics literature in demonstrating that, whenever a strong and valid, continuous is available, LIV might be preferable to conventional IV methods like 2SLS. Research paper 3 evaluates the performance of LIV and 2SLS according to mean bias and statistical efficiency in ATE and CATE estimates. While it was anticipated from theory that that 2SLS would not provide unbiased, efficient estimates of the ATE or CATE parameters under essential heterogeneity, the study also found that compared to 2SLS, LIV reported less biased, more efficient estimates of both sets of parameters, in settings with non-essential heterogeneity or homogeneity.

Two other key findings from research paper 3 are worth-emphasising. First, unlike 2SLS, LIV continues to perform well in settings with essential heterogeneity, but might require stronger IVs, or larger sample sizes to report small biases (<5%) and lower RMSE. Second, the study finds that 2SLS *can* be biased even in the setting where there is effect homogeneity if the instrument is not sufficiently strong (F<100). These findings align with previous work suggesting that without large sample sizes, if the IV is not sufficiently strong 2SLS can provide treatment effect estimates with substantial biases (Martens et al., 2006).

Research paper 3 also compares CATE estimates under 2SLS and LIV in three cohorts of the ESORT study, and reveals that the choice of the method matters in practice for the common setting in which the selection mechanism is unknown. While 2SLS fails to detect any signal of heterogeneity across subgroup estimates, LIV finds evidence that treatment effects of ES vary according to the patient's age, frailty level and number of comorbidities, which has direct relevance for policy-makers with respect to targeting scarce surgical resources.

6.4 Other general methodological contributions

In considering the application of the target trial framework to RWD, and evaluating the properties of IV methods across broader range of settings, including essential heterogeneity, the findings are relevant for the general causal inference and health econometrics literatures.

6.4.1 Insights from target trial relevant to the literature of observational epidemiology methods using Real-World Data in general and Instrumental Variable methods in particular

While the target trial framework has been previously applied in the epidemiological literature studies with rich information about treatment strategies, confounders and outcomes, there are few applications in the RWD setting that use routine (administrative) data. The resulting lack of guidance on how to address the challenges raised by RWD could cause bias and emulation failures to be inadvertently introduced into the studies (Franklin et al., 2020, 2021). Research paper 1 gives insights into how expert judgement might be used to make decisions about study design when the RWD is insufficient. While previous work has elaborated on methods for structured elicitation for HTA (Soares et al., 2018), this has not considered the potential relevance of formal or informal elicitation approaches for applying target trial emulation to RWD. Similarly, few studies have considered IV methods for emulating target trials (Danaei et al., 2018). Research paper 1 therefore contributes to this limited literature in describing how the design elements of RCTs can be emulated using LIV designs and in using expert opinion to define key standpoints of the study with respect to the population and comparator groups.

6.4.2 Evaluation of Instrumental Variable methods wider contexts

Most previous simulation work in the health econometrics literature evaluating IV methods ability to identify treatment effect parameters have been limited to settings where treatment effects had been assumed to be homogeneous (Ionescu-Ittu et al., 2012; Kang et al., 2015). The papers by Chapman and Brooks (2016) and Basu et al. (2018) did consider heterogenous treatment effects in evaluating the consistency of 2SLS and 2SRI in estimating the LATE and ATE parameters. However, none these studies considered LIV methods. Basu (2014) demonstrated the consistency of the LIV approach under ‘optimal’ conditions, which assumed large sample sizes and strong instruments, and allowed for essential heterogeneity, but that study did not formally test the performance of the method under scenarios such as weak identification or partial violation of the identification assumptions. Research paper 3 expands this previous work by Basu (2014) by considering a wider range of scenarios, defined by varying levels of IV strength and sample size, as well as across scenarios with different forms of treatment effect heterogeneity.

6.5 Limitations

In this section, I acknowledge general limitations relating to the unverifiability of the IV identification assumptions, and to other challenges that were not explored in this thesis. I then consider the interesting avenues for future research that this thesis provokes for methods for CEA using routine data.

6.5.1 Further challenges in applying target trial framework to Real-World Data

The list of recommendations outlined in research paper 1 does not aspire to offer solutions to all the issues that are raised in CEA that use routine data. Instead, it intends to offer practical recommendations for some common challenges that can arise with respect to confounding, when applying the target trial framework to assess the comparative effectiveness and cost-effectiveness from RWD settings in which individual participant data are available for all the comparators of interest. While the challenges identified in the ESORT study are common in RWD, as the analyst rarely has control over the treatment selection or data collection process, other complexities may well arise. For instance, there might be concerns around the accuracy of the outcome data. Information bias could emerge if for example information on a resource use measure, such as whether there were differences across the comparator groups in the way events, such as the receipt of surgery were recorded in the RWD (Rassen et al., 2021). The complex nature of RWD also means that, the recommendations outlined in research paper 1, will be need to be adapted to the specific context of the study. For instance, in studies that aim to evaluate complex treatment pathways, including subsequent treatment switching, it may be necessary to consider multiple definitions of time zero, and also forms of confounding, in particular time-varying confounding, beyond those considered in the ESORT study (section 6.6). Likewise in some settings, such as those where treatment switching occurs or there is non-adherence, might pose additional challenges for studies using RWD, as this might raise concerns about time-dependent confounding. Some papers have described how methods like IPW might be used for evaluating dynamic treatment regimens (Hernán et al., 2012), and recent work extending IV estimation with structural mean models in presence of time-varying confounding might be relevant (Shi et al., 2022).

Likewise, while the approach taken to elicitation was structured, it was pragmatic and used a modified Delphi approach to establish consensus across the panel of experts. Further work could more formally consider the uncertainty raised by divergent views across the expert panel (Soares et al., 2018).

6.5.2 Unverifiability of identification assumptions

Like any observational study, the CEA findings in research paper 2 rely to some extent on some unverifiable assumptions. The study describes the assumptions required for identification of marginal treatment effects with LIV, namely the exchangeability condition, the exclusion-restriction condition, the relevance assumption, and the monotonicity assumption. For example, it seems unlikely that there were imbalances in patients' prognosis across different levels of the TTO. For instance, the nature of the emergency setting, and the exclusion of patients who were referred to tertiary referral centres is likely to reduce the risk of bias due to the "doctor (hospital) shopping" phenomenon, which has been observed in other settings (Rassen et al., 2009). The study found that the TTO balanced the observed covariates, which gave some support to the requisite assumption that it was also able to achieve balance in the unmeasured characteristics.

One important challenge for the study was the exclusion-restriction, which requires that the IV only influences the outcome through its association with treatment assignment. To boost the plausibility of this assumption in the ESORT study, all analytical models included adjustment for proxies of the hospital's quality of acute care, for example through improved post-operative care, in addition to patient covariates (see chapter 4). Likewise, in defining the IV at the hospital level, instead of at the surgeon- or team-level, the study sought to minimise the risk of bias emerging from the association of the TTO with concomitant treatments, which is typically observed in preference-based IV settings (Baiocchi et al., 2021; Brookhart and Schneeweiss, 2007). While, after the adjustments for quality for care the exclusion-restriction was judged plausible, the study could have assessed this assumption more formally. For instance, with falsification tests, such as estimating whether ES had any effect on subgroups of always-treated patients to falsify this assumption[7] (Kang et al., 2013).

6.5.3 Comparators and metrics to evaluate the performance of Local Instrumental Variable methods in the simulation study

The simulation study in research paper 3 evaluated LIV and 2SLS across settings with different instruments, sample sizes and forms of treatment effect heterogeneity according to the mean bias in estimates and the rMSE. The findings expand the weak identification literature in finding that LIV has good estimation properties across a

[7] Note that the study assumes that all patients included in the study given the eligibility criteria have a level of their unmeasured covariates, such that they can be induced into treatment selection following a change in the level of the IV. This rules out the existence of patients who are never-takers or always-takers.

range of scenarios, and demonstrating that it can report estimates with less bias and lower RMSE than 2SLS (see section 6.6.2). However, while the study deliberately chose not to include methods that assume no unobserved confounding, as they lay beyond the scope this thesis, the comparison could have been extended to consider a broader range of comparators including 2SRI, which, unlike 2SLS, can in theory retrieve the ATE for the population in presence of unmeasured confounding (Basu et al., 2018; Terza et al., 2008). Also, the study sought to formally evaluate LIV methods in typical settings for the use of RWD according to levels of IV strength or sample size. However, I did not consider other issues that arise with IV approaches, including bias from violations of the exclusion-restriction assumption. Finally, the chosen metrics for the simulation study, bias and rMSE could have been supplemented by other measures of performance according to power or CI coverage probability (see section 6.6).

6.6 Areas for future research

This thesis identified a series of areas for further research.

6.6.1 Application of the principles of target trial framework and Local Instrumental Variables in settings with time-varying confounding

Future studies could expand on the methods described in the thesis to address related challenges. In particular, the ESORT study exemplifies the setting in which a single ‘one-off’ treatment is administered at a particular timepoint. Hence, further research is required to evaluate relative effectiveness in those settings in which sequences of treatment are provided across the time horizon of interest, which raises issues about time-varying confounding according to observed and unobserved factors. Methods like inverse probability of treatment weighting (IPW), and parametric g-computation have been proposed for estimating treatment effects of time-varying treatments in presence of observed confounders that are also time-dependent (see Daniel et al. (2013) for a review). However, the theory for estimating effects of time-varying treatments using IVs is far less developed. Recently, Tchetgen et al. (2018) considered IV estimation in the context of Marginal Structural Models, which were introduced by Robins (1997) for estimating joint effects time-varying treatment, but only in the context binary instruments. Further research could evaluate whether methods can be expanded to consider continuous IVs. This thesis did not consider other strategies for mitigating the risk of immortal time bias in studies using RWD and treatment strategies with a grace period. Methods like CCW described by Hernán and Robins (2016) could be adopted within IV designs in presence of time-varying confounding.

6.6.2 Extending simulation study design to consider additional metrics of performance for evaluating Local Instrumental Variable methods

While research paper 3 gives insights into the finite sample estimation properties of LIV methods according to different levels of IV strength, the simulation study did not evaluate the reliability of statistical inferences on treatment effects of these methods. It is well-known that, unlike inferences based on the weak-identification-robust Anderson-Rubin (AR) test statistic[8], t-ratio-based inferences estimator are subject to size/coverage distortions when instruments are weak (Dufour, 2003). Recent work by Lee et al. (2021) has quantified those distortions in the case of 2SLS, and suggested that applying a '95% confidence' requires that the first stage F statistic exceeds 100. However, it is unclear whether these findings apply to the continuous IV setting, where identification relies on the existence of a continuous IV that is not only valid, but also sufficiently strong to ensure that there is a level of the IV at which all the individuals in the sample 'comply' (i.e., are shifted or selected into treatment) (Basu et al., 2007).

The strength of the instrument and the sample size have also been shown to influence the study's power to detect true causal effects (minimise the risk of type-II errors) of IV methods. Recently, Keane and Neal (2021) and Angrist and Kolesár (2021) showed that the power conveyed by conventional F-statistic values (i.e., around 10) with 2SLS can be low. While, this is likely to be the case for LIV too, power evaluations have yet to be extended to continuous IV. Therefore, future research could extend the simulation study design in research paper 3 to formally evaluate the reliability of inference with LIV by considering coverage of the 95% CI, and power to detect a causal effect based on the CI. Statistical power could also be affected by imbalances in treatment assignment in the case of LIV, just as with 2SLS (Campbell and Gustafson, 2018; Myers et al., 2011). Therefore, the simulation study design could also be expanded to consider scenarios where the proportion of patients exposed (assigned) to treatment is varied alongside the instrument strength and sample size.

[8] The test has correct size regardless of sample size and strength of the IV.

6.6.3 Use of data-adaptive methods alongside Local Instrumental Variable methods

Consistent estimation of treatment effect parameters using the methods described in the thesis requires that the parametric models for the outcome and treatment selection are correctly specified. This can be a challenging task, especially in settings with high-dimensional data, that carries a high risk of bias due to model misspecification. Machine learning methods have already been applied for high-dimensional covariate selection in IV designs using 2SLS methods for policy evaluation (Bakx et al., 2020; Martin et al., 2022). Similar approaches could be adopted in the continuous/multi-valued IV setting. Recently, a doubly robust (DR) estimator of the MTE curve was developed by Kennedy et al. (2019). The DR estimator, together with the models for the outcome and treatment, requires the specification and estimation of an additional nuisance function, which models how the instrument depends on the covariates. In this thesis, following (Heckman and Vytlacil, 1999, 2001), the estimation of target treatment effect parameters was done defining fully parametric models. Instead, Kennedy et al. (2019)'s proposed approach relaxes the assumption that both parametric models are correctly specified. Another advantage of this flexible method is that the target estimand can be made conditional on covariates of interest, and not the full covariate space. Further work could build on this work using data-adaptive covariate selection methods such as least absolute shrinkage and selection operator (LASSO) for confounding adjustment in high dimensional settings (Belloni et al., 2014).

6.7 Recommendations for applied researchers

6.7.1 Apply the set of recommendations for target trial using Real-World Data to inform Health Technology Assessment provided in research paper 1

The target trial emulation framework should be adopted as part of CEA designs that use RWD. Researchers should carefully assess the sufficiency and adequacy of the data to answer the research question, and in particular the challenges raised in research paper 1. When the RWD is insufficient or inadequate to emulate the design elements of the target trial, researchers should consider the recommendations provided in research paper 1. These should be adopted alongside published questionnaires and checklists for economic evaluations, and other available tools for evaluating the plausibility of underlying assumptions of statistical methods (Faria et al., 2015; Kreif et al., 2013).

6.7.2 When the study has access to a continuous or multi-valued instrument, consider using Local Instrumental Variable methods for informing treatment effects as described in research paper 2

The thesis describes novel methods for estimating treatment effects in the presence of unmeasured confounding and heterogeneity. It is recommended that in settings where essential heterogeneity is anticipated, researchers consider LIV methods to report robust estimates of treatment effects. More generally, LIV methods should be considered in any setting in which there is interest in evaluating heterogeneous treatment effects. Methods described in research paper 2 and 3 could be used to report estimates at the individual-level, or aggregated over subgroups of interest.

6.7.3 Consider whether the strength of the instrument is enough to ensure low bias and sufficient statistical efficiency given the available sample size and form of treatment effect heterogeneity as described in research paper 3

This thesis recommends that studies using LIV methods carefully consider the relevance assumption. The strength of the instrument, as measured by the value of the first stage F statistic, should not be judged according to whether it exceeds a fixed threshold. Whether the IV is sufficiently strong should be assessed alongside the sample size available, and the interaction between confounding and heterogeneity. For example, moderate and small samples may require a much stronger IV, i.e., higher F statistic, whereas if the sample size is very large, or if the treatment effects are expected to be homogeneous across individuals, then a weaker IV may suffice.

6.8 Implications for policy-making

Real-world evidence is being increasingly adopted by HTA agencies to inform decision-making in healthcare. The recently-published, NICE methods guidance highlights the need for real-world evidence, but acknowledges the lack of trust in this type of evidence as a major barrier for its adoption (NICE, 2022). The target trial framework offers a means to increase trust in this type of evidence from observational studies that use RWD, and improve the use of RWD in HTA. The target trial framework can help decision-makers and clinical experts to judge the rigour and reliability of the evidence according to how closely the study replicates (emulates) the design elements of the target trial. The recommendations outlined in research paper 1 could improve the process of critical appraisal of CEA that use routine data in HTA evaluations, and help those using the evidence to judge the appropriateness of the study design. For example, the risk of selection bias in the CEA could be judged according to whether the definition of time (day zero or baseline) is aligned with the time that eligibility is assessed.

The ESORT study addressed some of the gaps in evidence evaluating the cost-effectiveness of ES. In particular, the LIV approach found that while there is no indication that either ES or NES are cost-effective overall, there are subgroups of patients for which further uptake of NES strategies could result in better outcomes and cost, albeit with uncertainty about the magnitude of these effects reflecting the respective sample sizes. The LIV approach identified subgroups for which there may be cost savings from the increased uptake of ES or NES strategies. It found that NES strategies are more cost-effective for patients with moderate or severe frailty (acute appendicitis, abdominal wall hernia), and with at least two comorbidities (hernia), or in older age groups (appendicitis). Likewise, NES was the more cost-effective alternative for patients with diverticular disease with perforation and abscess. These findings suggest that currently NES (ES) might be (over)underused in some subgroups of patients with these conditions, and that redirecting resources could result in gains in terms of efficiency. The observed patterns of treatment effect heterogeneity across population subgroups could help inform future guidance for emergency admissions about triage for ES according to baseline risk profiles. However, further evidence is required to inform how best to ‘personalise’ the choice of strategy for patients for these conditions. These findings could help target future trials to provide more granular evidence about which prognostic factors play a role in explaining heterogeneous responses to ES.

6.9 Conclusions

This thesis aimed to address the lack of guidance for designing CEA that use routinely-collected individual-level RWD in the presence of unmeasured confounding and heterogeneity. The thesis carefully examined the challenges that could arise in extending the target trial framework to CEA that use RWD, and offers a series of recommendations for future studies.

The risk of unmeasured confounding is arguably the biggest threat to providing reliable evidence from CEA that use RWD. This thesis extends LIV methods for tackling the risk of residual confounding and heterogeneity from unmeasured covariates to a CEA that uses individual-level routine data from England. I formally assess the main requirements for the LIV approach and provide guidelines for future studies. While these assumptions must also be made assessable and carefully assessed, these methods have the potential to use RWD to inform the personalisation of treatment choice. The focus of the thesis is on CEA that intend to inform decision-making within the HTA context, but the principles developed can improve how RWD is used in comparative effectiveness and cost-effectiveness studies more generally.

References

Angrist J and Kolesár M (2021) One Instrument to Rule Them All: The Bias and Coverage of Just-Id IV. *SSRN Electronic Journal*. DOI: 10.2139/ssrn.3953944.

Bakx P, Wouterse B, van Doorslaer E, et al. (2020) Better off at home? Effects of nursing home eligibility on costs, hospitalizations and survival. *Journal of Health Economics* 73. Elsevier B.V. DOI: 10.1016/j.jhealeco.2020.102354.

Basu A (2014) Estimating person-centered treatment (PeT) effects using instrumental variables: an application to evaluating prostate cancer treatments. *JOURNAL OF APPLIED ECONOMETRICS* 29: 671–691. DOI: 10.1002/jae.

Basu A, Heckman JJ, Navarro-Lozano S, et al. (2007) Use of instrumental variables in the presence of heterogeneity and self-selection: an application to treatments of breast cancer patients. *Health Economics* 16(2007): 1133–1157. DOI: 10.1002/hec.1291.

Basu A, Coe NB and Chapman CG (2018) 2SLS versus 2SRI: Appropriate methods for rare outcomes and/or rare exposures. *Health Economics* 27(6): 937–955. DOI: 10.1002/hec.3647.

Baiocchi M, Cheng J and Small DS (2021) Tutorial in Biostatistics: Instrumental Variable Methods for Causal Inference. *Statistics in Medicine*: 1–55. DOI: 10.1002/sim.0000.

Belloni A, Chernozhukov V and Hansen C (2014) High-Dimensional Methods and Inference on Structural and Treatment Effects. *Journal of Economic Perspectives* 28(2): 29–50. DOI: 10.1257/jep.28.2.29.

Brookhart MA and Schneeweiss S (2007) Preference-based instrumental variable methods for the estimation of treatment effects: Assessing validity and interpreting results. *International Journal of Biostatistics* 3(1): 1–19. DOI: 10.2202/1557-4679.1072.

Campbell H and Gustafson P (2018) The Validity and Efficiency of Hypothesis Testing in Observational Studies with Time-Varying Exposures. *Observational Studies* 4(1): 260–291. DOI: 10.1353/obs.2018.0010.

Chapman CG and Brooks JM (2016) Treatment Effect Estimation Using Nonlinear Two-Stage Instrumental Variable Estimators: Another Cautionary Note. *Health Services Research* 51(6). Blackwell Publishing Inc.: 2375–2394. DOI: 10.1111/1475-6773.12463.

Danaei G, García Rodríguez LA, Cantero OF, et al. (2018) Electronic medical records can be used to emulate target trials of sustained treatment strategies. *Journal of Clinical Epidemiology* 96. DOI: 10.1016/j.jclinepi.2017.11.021.

Daniel RM, Cousens SN, De Stavola BL, et al. (2013) Methods for dealing with time-dependent confounding. *Statistics in medicine* 32(9). Wiley Online Library: 1584–1618.

DiazOrdaz K and Grieve R (2019) Noncompliance and Missing Data in Health Economic Evaluation. Oxford University Press. DOI: 10.1093/acrefore/9780190625979.013.94.

Drummond MF, Sculpher MJ, Torrance GW, et al. (2005) *Methods for the Economic Evaluation of Health Care Programmes*. Oxford medical publications. Oxford University Press. Available at: https://books.google.es/books?id=CxWzQgAACAAJ.

Dufour J (2003) Identification, weak instruments, and statistical inference in econometrics. *Canadian Journal of Economics/Revue canadienne d'économique* 36(4). Wiley Online Library: 767–808.

Faria R, Hernández Alava M, Manca A, et al. (2015) *NICE DSU Technical Support Document 17: The use of observational data to inform estimates of treatment effectiveness in technology appraisal: Methods for comparative individual patient data.*

Flum DR, Davidson GH, Monsell SE, et al. (2020) A Randomized Trial Comparing Antibiotics with Appendectomy for Appendicitis. *New England Journal of Medicine* 383(20): 1907–1919. DOI: 10.1056/nejmoa2014320.

Franklin JM, Glynn RJ, Suissa S, et al. (2020) Emulation Differences vs. Biases When Calibrating Real-World Evidence Findings Against Randomized Controlled Trials. *Clinical Pharmacology and Therapeutics* 107(4): 735–737. DOI: 10.1002/cpt.1793.

Franklin JM, Patorno E, Desai RJ, et al. (2021) Emulating Randomized Clinical Trials with Nonrandomized Real-World Evidence Studies: First Results from the RCT DUPLICATE Initiative. *Circulation* (143): 1002–1013. DOI: 10.1161/CIRCULATIONAHA.120.051718.

Heckman JJ and Vytlacil EJ (1999) Local instrumental variables and latent variable models for identifying and bounding treatment effects. *Proceedings of the National Academy of Sciences of the United States of America* 96: 4730–4734. DOI: 10.1073/pnas.96.8.4730.

Heckman JJ and Vytlacil EJ (2001) Policy-Relevant Treatment Effects. *American Economic Review* 91(2): 107–111. DOI: 10.1257/aer.91.2.107.

Hernán MA (2004) A definition of causal effect for epidemiological research. *Journal of Epidemiology and Community Health* 58(4): 265–271. DOI: 10.1136/jech.2002.006361.

Hernán, M.A., Lanoy, E., Costagliola, D. and Robins, J.M. (2006), Comparison of Dynamic Treatment Regimes via Inverse Probability Weighting. Basic & Clinical Pharmacology & Toxicology, 98: 237-242. https://doi.org/10.1111/j.1742-7843.2006.pto_329.x

Hernán MA and Taubman SL (2008) Does obesity shorten life? The importance of well-defined interventions to answer causal questions. *International Journal of Obesity* 32: S8–S14. DOI: 10.1038/ijo.2008.82.

Hernán MA and Robins JM (2016) Using Big Data to Emulate a Target Trial When a Randomized Trial Is Not Available. *American Journal of Epidemiology* 183(8). Oxford University Press: 758–764. DOI: 10.1093/aje/kwv254.

Hernán MA, Sauer BC, Hernández-Díaz S, et al. (2016) Specifying a target trial prevents immortal time bias and other self-inflicted injuries in observational analyses. *Journal of Clinical Epidemiology* 79: 70–75. DOI: 10.1016/j.jclinepi.2016.04.014.Specifying.

Husereau D, Drummond M, Petrou S, et al. (2013) Consolidated health economic evaluation reporting standards (CHEERS)-explanation and elaboration: A report of the ISPOR health economic evaluation publication guidelines good reporting practices task force. *Value in Health* 16(2). Elsevier: 231–250. DOI: 10.1016/j.jval.2013.02.002.

Ionescu-Ittu R, Abrahamowicz M and Pilote L (2012) Treatment effect estimates varied depending on the definition of the provider prescribing preference-based instrumental variables. *Journal of Clinical Epidemiology* 65(2). Elsevier Inc: 155–162. DOI: 10.1016/j.jclinepi.2011.06.012.

Javanmard-Emamghissi H, Hollyman M, Boyd-Carson H, et al. (2021) Antibiotics as first-line alternative to appendicectomy in adult appendicitis: 90-day follow-up from a prospective, multicentre cohort study. *British Journal of Surgery*: 1–9. DOI: 10.1093/bjs/znab287.

Kang H, Kreuels B, Adjei O, et al. (2013) The causal effect of malaria on stunting: a Mendelian randomization and matching approach. *International journal of epidemiology* 42(5). Oxford University Press: 1390–1398.

Kang H, Cai TT and Small DS (2015) A simple and robust confidence interval for causal effects with possibly invalid instruments. *arXiv*: 1–15. Available at: http://arxiv.org/abs/1504.03718 (accessed 9 November 2021).

Keane M and Neal T (2021) *A Practical Guide to Weak Instruments. UNSW Economics Working Paper No. 2021-05d.*

Kennedy EH, Lorch S and Small DS (2019) Robust causal inference with continuous instruments using the local instrumental variable curve. *Journal of the Royal Statistical Society. Series B: Statistical Methodology* 81(1): 121–143. DOI: 10.1111/rssb.12300.

Kreif N, Grieve R and Sadique MZ (2013) Statistical methods for cost-effectiveness analyses that use observational data: a critical appraisal tool and review of current practice. *Health Economics* 22: 486–500. DOI: 10.1002/hec.

Kreif N, Mirelman A, Moreno Serra R, et al. (2020) Who benefits from health insurance?: Uncovering heterogeneous policy impacts using causal machine learning. Centre for Health Economics, University of York.

Latimer NR, Abrams KR, Lambert PC, et al. (2014) Adjusting survival time estimates to account for treatment switching in randomized controlled trials—an economic evaluation context: methods, limitations, and recommendations. *Medical Decision Making* 34(3). Sage Publications Sage CA: Los Angeles, CA: 387–402.

Lee D, McCrary J, Moreira MJ, et al. (2021) *Valid T-Ratio Inference for IV. National Bureau of Economic Research Working Paper Series (No. w29124).* DOI: 10.2139/ssrn.3901588.

Lodi S, Phillips A, Lundgren J, et al. (2019) Effect Estimates in Randomized Trials and Observational Studies: Comparing Apples With Apples. *American Journal of Epidemiology* 188(8). Oxford Academic: 1569–1577. DOI: 10.1093/AJE/KWZ100.

Martens EP, Pestman WR, De Boer A, et al. (2006) Instrumental variables: Application and limitations. *Epidemiology* 17(3): 260–267. DOI: 10.1097/01.ede.0000215160.88317.cb.

Martin S, Claxton K, Lomas J, et al. (2022) How Responsive is Mortality to Locally Administered Healthcare Expenditure? Estimates for England for 2014/15. *Applied Health Economics and Health Policy*. Springer: 1–16.

Michael H, Cui Y and Tchetgen EJT (n.d.) An Inverse Probability Weighting Approach for Instrumental Variable Estimation of Marginal Structural Mean Models.

Myers JA, Rassen JA, Gagne JJ, et al. (2011) Effects of adjusting for instrumental variables on bias and precision of effect estimates. *American Journal of Epidemiology* 174(11): 1213–1222. DOI: 10.1093/aje/kwr364.

National Institute for Health and Care Excellence (2022) *NICE real-world evidence framework*. London.

Nixon RM and Thompson SG (2005) Methods for incorporating covariate adjustment, subgroup analysis and between-centre differences into cost-effectiveness evaluations. *Health Economics* 14(12): 1217–1229. DOI: 10.1002/hec.1008.

O'Leary DP, Walsh SM, Bolger J, et al. (2021) A Randomized Clinical Trial Evaluating the Efficacy and Quality of Life of Antibiotic-only Treatment of Acute Uncomplicated Appendicitis: Results of the COMMA Trial. *Annals of surgery* 274(2): 240–247. DOI: 10.1097/SLA.0000000000004785.

Patorno, E, Schneeweiss S, and Wang S. (2020) Transparency in Real-World Evidence (RWE) Studies to Build Confidence for Decision-Making: Reporting RWE Research in Diabetes. *Diabetes, Obesity and Metabolism* 22 (S3): 45–59. https://doi.org/10.1111/dom.13918.

Philips Z, Ginnelly L, Sculpher M, et al. (2004) Review of guidelines for good practice in decision-analytic modelling in health technology assessment. *Health Technology Assessment* 8(36). DOI: 10.3310/hta8360.

Polsky D and Basu A (2012) *Chapter 46: Selection Bias in Observational Data.* (AM Jonesed.). he Elgar C. Edward Elgar Publishing.

Rassen JA, Brookhart MA, Glynn RJ, et al. (2009) Instrumental variables I: instrumental variables exploit natural variation in nonexperimental data to estimate causal relationships. *Journal of Clinical Epidemiology* 62(12): 1226–1232. DOI: 10.1016/j.jclinepi.2008.12.005.Instrumental.

Rassen JA, Murk W and Schneeweiss S (2021) Real-world evidence of bariatric surgery and cardiovascular benefits using electronic health records data: A lesson in bias. *Diabetes, Obesity and Metabolism* 23(7): 1453–1462. DOI: 10.1111/dom.14338.

Robins JM (1997) Causal Inference from Complex Longitudinal Data. In: *Latent Variable Modeling and Applications to Causality.* New York, NY: Springer Berlin Heidelberg, pp. 69–117. DOI: 10.1007/978-1-4612-1842-5_4.

Sadique Z, Grieve R, Diaz-Ordaz K, et al. (2022) A Machine-Learning Approach for Estimating Subgroup- and Individual-Level Treatment Effects: An Illustration Using the 65 Trial. *Medical Decision Making* 42(7): 923–936. DOI: 10.1177/0272989X221100717.

Sekhon JS and Grieve R (2012) A matching method for improving covariate balance in cost-effectiveness analyses. *Health Economics* (21): 695–713. DOI: 10.1002/hec.1748.

Shi, J., Swanson, S. A., Kraft, P., Rosner, B., De Vivo, I., & Hernán, M. A. (2021). Instrumental variable estimation for a time-varying treatment and a time-to-event outcome via structural nested cumulative failure time models. BMC medical research methodology, 21(1), 258. https://doi.org/10.1186/s12874-021-01449-w

Skivington K, Matthews L, Simpson SA, et al. (2021) Framework for the development and evaluation of complex interventions: Gap analysis, workshop and consultation-informed update. *Health Technology Assessment* 25(57). DOI: 10.3310/HTA25570.

Soares MO, Sharples L, Morton A, et al. (2018) Experiences of structured elicitation for model-based cost-effectiveness analyses. *Value in health* 21(6). Elsevier: 715–723.

Swanson SA (2017) Instrumental Variable Analyses in Pharmacoepidemiology: What Target Trials Do We Emulate? *Current Epidemiology Reports* 4(4). Current Epidemiology Reports: 281–287. DOI: 10.1007/s40471-017-0120-1.

Tchetgen EJT, Michael H and Cui Y (2018) Marginal structural models for time-varying endogenous treatments: A time-varying instrumental variable approach. *arXiv*: 1–27.y

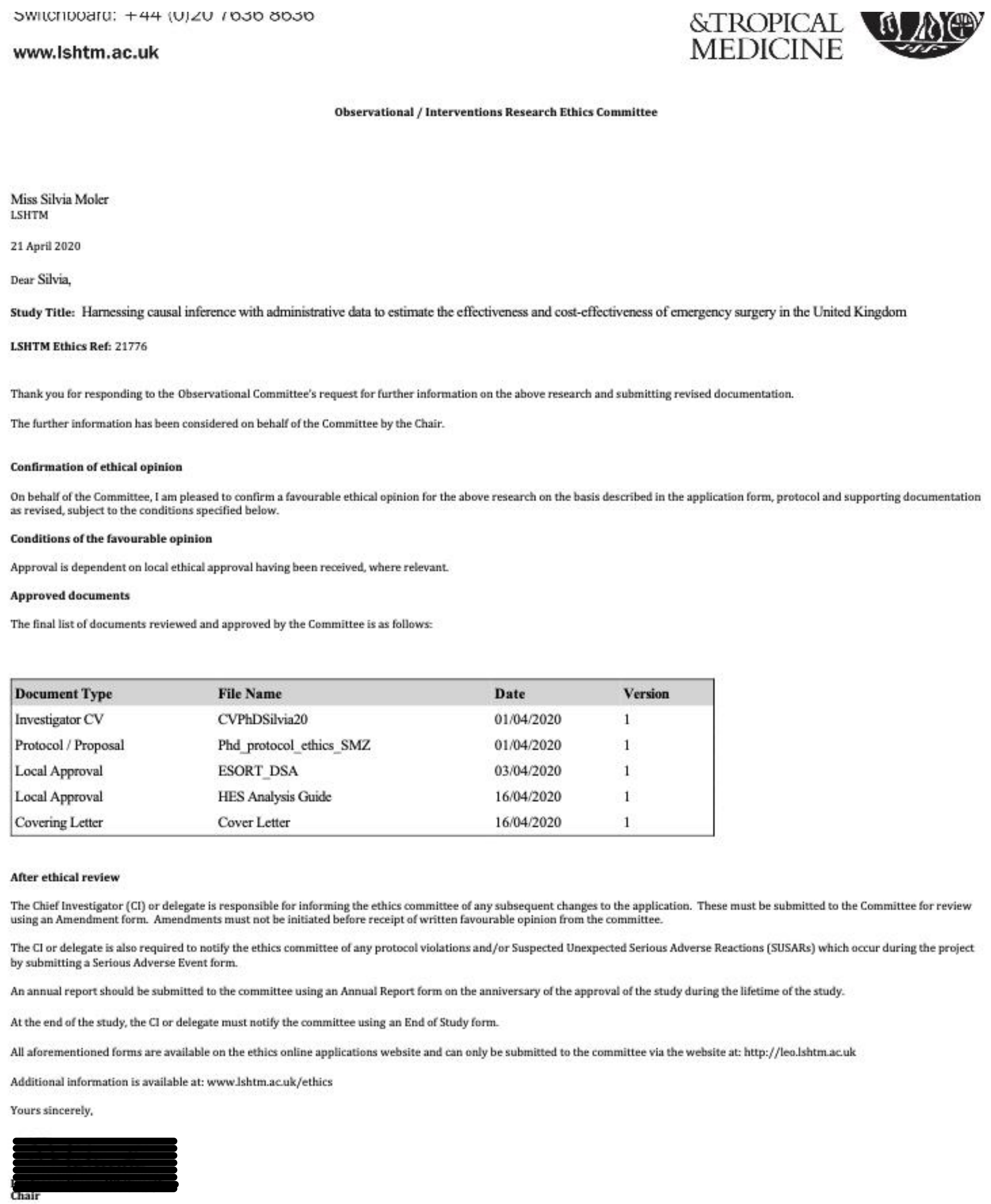
Terza J V., Basu A and Rathouz PJ (2008) Two-stage residual inclusion estimation: Addressing endogeneity in health econometric modeling. *Journal of Health Economics* 27(3). NIH Public Access: 531–543. DOI: 10.1016/j.jhealeco.2007.09.009.

Walker A, Patrick AR, Lauer MS, et al. (2013) A tool for assessing the feasibility of comparative effectiveness research. *Comparative Effectiveness Research*: 11. DOI: 10.2147/cer.s40357

Willan AR and Briggs AH (2006) *Statistical Analysis of Cost-Effectiveness Data.* John Wiley & Sons.

Appendices

Appendix A. Ethics approval



Switchboard: +44 (0)20 7636 8636

www.lshtm.ac.uk

&TROPICAL
MEDICINE

Observational / Interventions Research Ethics Committee

Miss Silvia Moler
LSHTM

21 April 2020

Dear Silvia,

Study Title: Harnessing causal inference with administrative data to estimate the effectiveness and cost-effectiveness of emergency surgery in the United Kingdom

LSHTM Ethics Ref: 21776

Thank you for responding to the Observational Committee's request for further information on the above research and submitting revised documentation.

The further information has been considered on behalf of the Committee by the Chair.

Confirmation of ethical opinion

On behalf of the Committee, I am pleased to confirm a favourable ethical opinion for the above research on the basis described in the application form, protocol and supporting documentation as revised, subject to the conditions specified below.

Conditions of the favourable opinion

Approval is dependent on local ethical approval having been received, where relevant.

Approved documents

The final list of documents reviewed and approved by the Committee is as follows:

Document Type	File Name	Date	Version
Investigator CV	CVPhDSilvia20	01/04/2020	1
Protocol / Proposal	Phd_protocol_ethics_SMZ	01/04/2020	1
Local Approval	ESORT_DSA	03/04/2020	1
Local Approval	HES Analysis Guide	16/04/2020	1
Covering Letter	Cover Letter	16/04/2020	1

After ethical review

The Chief Investigator (CI) or delegate is responsible for informing the ethics committee of any subsequent changes to the application. These must be submitted to the Committee for review using an Amendment form. Amendments must not be initiated before receipt of written favourable opinion from the committee.

The CI or delegate is also required to notify the ethics committee of any protocol violations and/or Suspected Unexpected Serious Adverse Reactions (SUSARs) which occur during the project by submitting a Serious Adverse Event form.

An annual report should be submitted to the committee using an Annual Report form on the anniversary of the approval of the study during the lifetime of the study.

At the end of the study, the CI or delegate must notify the committee using an End of Study form.

All aforementioned forms are available on the ethics online applications website and can only be submitted to the committee via the website at: http://leo.lshtm.ac.uk

Additional information is available at: www.lshtm.ac.uk/ethics

Yours sincerely,

[redacted]

Chair

Appendix B. Chapter 3

Appendix B.1. Definition of criteria for inclusion into acute appendicitis and acute gallstone disease cohorts based on expert clinical opinion

Acute appendicitis		**Acute gallstone disease**	
HES ICD-10 diagnostic codes	**Did the clinical panel favour the inclusion in definition of the condition? (% favourable)**	**HES ICD-10 diagnostic codes**	**Did the clinical panel favour the inclusion in definition of the condition? (% favourable)**
Acute appendicitis (K35)	Yes (100)	Calculus of gallbladder with acute cholecystitis (K80.0)	Yes (100)
Acute appendicitis with generalized peritonitis (K35.2)	Yes (83)	Calculus of gallbladder with other cholecystitis (K80.1)	Yes (100)
Acute appendicitis with localized peritonitis (K35.3)	Yes (100)	Calculus of gallbladder without cholecystitis (K80.2)	Yes (83)
Acute appendicitis, other and unspecified (K35.8)	Yes (100)	Calculus of bile duct with cholangitis (K80.3)	No (42)
Unspecified appendicitis (K37)	Yes (75)	Calculus of bile duct with cholecystitis (K80.4)	No (58)
		Calculus of bile duct without cholangitis or cholecystitis (K80.5)	No (42)
		Other cholelithiasis (K80.8)	No (50)

*List of HES ICD-10 diagnosis codes assessed for inclusion in definition of Acute Appendicitis and Acute gallstone disease in ESORT study's target trial. 12 experts in the clinical panel were consulted in a two-round Delphi process. Panellists had the opportunity to discuss the results of the first round before providing their responses in the second round. A diagnostic code was included if at least 75% panellists favoured its inclusion in the second round. The second column presents the results of the second round. HES: Hospital Episode Statistics; ICD: International Classification of Diseases.

Appendix B.2. Definition of Emergency Surgery for acute appendicitis and acute gallstone disease based on expert clinical opinion

Acute appendicitis		Acute gallstone disease	
OPCS code	Did the clinical panel favour the inclusion in definition of ES? (% favourable)	OPCS code	Did the clinical panel favour the inclusion in definition of ES? (% favourable)
Ileectomy and anastomosis of ileum to ileum (G693)	No (0)	Total cholecystectomy and excision of surrounding tissue (J181)	Yes (100)
Ileectomy and anastomosis of ileum to colon (G694)	No (50)	Total cholecystectomy and exploration of common bile duct (J182)	Yes (100)
Unspecified excision of ileum (G699)	No (0)	Total cholecystectomy (J183)	Yes (100)
Emergency excision of abnormal appendix and drainage (H011)	Yes (100)	Partial cholecystectomy and exploration of common bile duct (J184)	Yes (100)
Emergency excision of abnormal appendix (JH012)	Yes (100)	Partial cholecystectomy (J185)	Yes (100)
Emergency excision of normal appendix (H013)	Yes (83)	Other specified excision of gall bladder (J188)	Yes (92)
Other specified emergency excision of appendix (H018)	Yes (100)	Unspecified excision of gall bladder (J189)	Yes (92)
Unspecified emergency excision of appendix (H019)	Yes (100)	Open removal of calculus from gall bladder (J211)	Yes (75)
Interval appendicectomy (H021)	No (0)	Drainage of gall bladder (J212)	Yes (100)
Planned delayed appendicectomy NEC (H022)	No (0)	Drainage of tissue surrounding gall bladder (J213)	Yes (75)
Prophylactic appendicectomy NEC (H023)	No (0)	Percutaneous drainage of gall bladder (J241)	Yes (58)
Incidental appendicectomy (H024)	No (0)	Open removal of calculus from bile duct (J332)	No (8)

Acute appendicitis		**Acute gallstone disease**	
OPCS code	**Did the clinical panel favour the inclusion in definition of ES? (% favourable)**	**OPCS code**	**Did the clinical panel favour the inclusion in definition of ES? (% favourable)**
Other specified other excision of appendix (H028)	No (8)	Drainage of bile duct (J333)	No (8)
Unspecified other excision of appendix (H029)	No (42)	Sphincteroplasty of bile duct using duodenal approach (J342)	No (0)
Drainage of abscess of appendix (H031)	Yes (100)	Unspecified plastic repair of sphincter of Oddi using duodenal approach (J349)	No (0)
Drainage of appendix (H032)	Yes (100)	Sphincterotomy of bile duct using duodenal approach (J352)	No (0)
Exteriorisation of appendix (H033)	No (0)	Unspecified incision of sphincter of Oddi using duodenal approach (J359)	No (0)
Other specified other operations on appendix (H038)	No (8)	Operative cholangiography through cystic duct (J372)	No (17)
Extended right hemicolectomy and anastomosis of ileum to colon (H062)	Yes (75)	Direct puncture operative cholangiography (J373)	No (8)
Right hemicolectomy and end to end anastomosis of ileum to colon (H071)	Yes (100)	Endoscopic sphincterotomy of sphincter of Oddi and removal of calculus (J381)	No (17)
Right hemicolectomy and side to side anastomosis of ileum to transverse colon (H072)	Yes (100)	Endoscopic sphincterotomy of sphincter of Oddi and insertion of tubal prosthesis into bile duct (J382)	No (17)
Right hemicolectomy and anastomosis (H073)	Yes (100)	Other specified endoscopic incision of sphincter of Oddi (J388)	No (8)

Acute appendicitis		Acute gallstone disease	
OPCS code	**Did the clinical panel favour the inclusion in definition of ES? (% favourable)**	**OPCS code**	**Did the clinical panel favour the inclusion in definition of ES? (% favourable)**
Right hemicolectomy and ileostomy (H074)	Yes (100)	Unspecified endoscopic incision of sphincter of Oddi (J389)	No (8)
Other specified other excision of right hemicolon (H078)	Yes (75)	Endoscopic sphincterotomy of accessory ampulla of Vater (J391)	No (0)
Unspecified other excision of right hemicolon (H079)	Yes (67)	Endoscopic retrograde insertion of tubal prosthesis into both hepatic ducts (J401)	No (0)
Unspecified opening of abdomen (T309)	No (8)	Endoscopic retrograde insertion of tubal prosthesis into bile duct (J402)	No (0)
Open drainage of pelvic abscess (T342)	Yes (100)	Endoscopic retrograde renewal of tubal prosthesis in bile duct (J403)	No (0)
Open drainage of abdominal abscess (T343)	Yes (100)	Endoscopic retrograde removal of tubal prosthesis from bile duct (J404)	No (0)
Image controlled percutaneous drainage of pelvic abscess (T452)	No (50)	Endoscopic retrograde insertion of expanding covered metal stent into bile duct (J405)	No (0)
Image controlled percutaneous drainage of abdominal abscess (T453)	No (50)	Endoscopic retrograde insertion of expanding metal stent into bile duct (J406)	No (0)
Image controlled percutaneous drainage of lesion of abdominal cavity (T454)	No (42)	Endoscopic retrograde extraction of calculus from bile duct (J411)	No (8)
Irrigation of peritoneal cavity (T463)	Yes (83)	Endoscopic dilation of bile duct (J412)	No (0)

Acute appendicitis		Acute gallstone disease	
OPCS code	**Did the clinical panel favour the inclusion in definition of ES? (% favourable)**	**OPCS code**	**Did the clinical panel favour the inclusion in definition of ES? (% favourable)**
Other specified other drainage of peritoneal cavity (T468)	Yes (75)	Endoscopic retrograde lithotripsy of calculus of bile duct (J413)	No (8)
		Other specified other therapeutic endoscopic retrograde operations on bile duct (J418)	No (0)
		Endoscopic retrograde insertion of tubal prosthesis into pancreatic duct (J421)	No (0)
		Endoscopic retrograde cholangiopancreatography and biopsy of lesion of ampulla of Vater (J431)	No (0)
		Other specified diagnostic endoscopic retrograde examination of bile duct and pancreatic duct (J438)	No (0)
		Unspecified diagnostic endoscopic retrograde examination of bile duct and pancreatic duct (J439)	No (0)
		Endoscopic retrograde cholangiography and biopsy of lesion of bile duct (J441)	No (0)
		Unspecified diagnostic endoscopic retrograde examination of bile duct (J449)	No (0)
		Unspecified diagnostic endoscopic retrograde examination of pancreatic duct (J459)	No (0)

Acute appendicitis		Acute gallstone disease	
OPCS code	**Did the clinical panel favour the inclusion in definition of ES? (% favourable)**	**OPCS code**	**Did the clinical panel favour the inclusion in definition of ES? (% favourable)**
		Percutaneous insertion of tubal prosthesis into common bile duct (J475)	No (0)
		Percutaneous transhepatic biliary drainage single (J486)	No (0)
		T tube cholangiography (J501)	No (8)
		Percutaneous cholangiography (J502)	No (0)
		Percutaneous transhepatic cholangiography (J505)	No (0)
		Unspecified endoscopic ultrasound examination of bile duct (J539)	No (0)
		Unspecified diagnostic endoscopic examination of peritoneum (T439)	No (0)

* List of Hospital Episode Statistics (HES) Office of Population Censuses and Surveys (OPCS) procedures assessed for inclusion in definition of emergency surgery (ES) for acute appendicitis and acute gallstone disease. 12 experts in the clinical panel were consulted in a two-round Delphi process. Panellists had the opportunity to discuss the results of the first round before providing their responses in the second round. A procedure code was included if at least 50% of the panellists (6) favoured its inclusion in the second round. The third column presents the results of the second round.

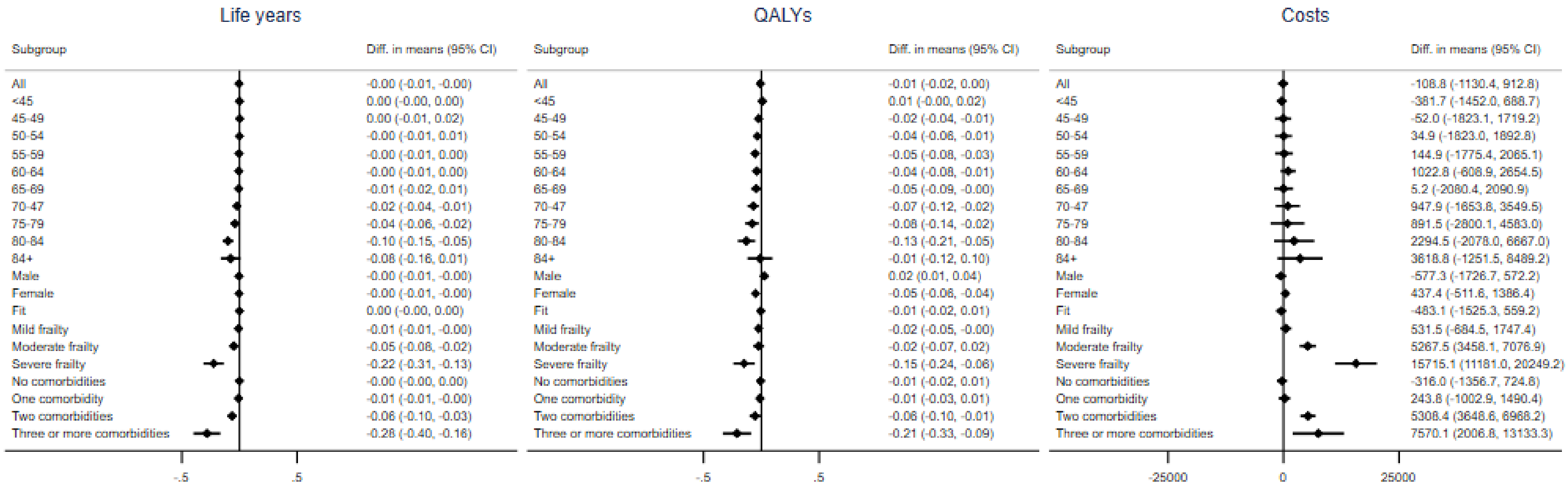
Appendix B.3. Estimated incremental costs (£GBP 2019/20) at one year of emergency surgery vs non-emergency surgery strategies with unadjusted differences, regression adjustment and the local instrumental variable (LIV) approach

Mean differences (95% CI)	Unadjusted differences	GLM regression	LIV approach
		Acute appendicitis (N=268,144)	
Costs	-413 (-514, -313)	266 (177, 354)	-109 (-1,130, 913)
Life years	0.014 (0.012, 0.016)	0.005 (0.004, 0.007)	-0.003 (-0.006, -0.001)
QALYs	0.051 (0.046, 0.056)	0.008 (0.007, 0.010)	-0.009 (-0.022, 0.003)
Net benefit	1,431 (1259, 1603)	-223 (-342, -104)	-86.2 (-1,163, 991)
		Acute gallstone disease (N=240,977)	
Costs	-251 (-386, -115)	281 (170, 393)	-76.8 (-702, 548)
Life years	0.013 (0.011, 0.014)	0.003 (0.002, 0.004)	-0.009 (-0.022, 0.005)
QALYs	0.038 (0.035, 0.040)	0.005 (0.004, 0.006)	0.007 (-0.001, 0.015)
Net benefit	1,002 (832, 1171)	-220 (-316, 124)	221 (-450, 892)

Variables used for adjustment in models: age (years), sex, ethnicity, index of multiple deprivation (quintiles), number of comorbidities (Charlson index), frailty level (SCARF index), method of admission, year fixed effects, proxies for the quality of acute care within the hospital. CI: confidence interval; QALYs: quality-adjusted life year.

Appendix B.4. Forest plots of estimated incremental life years (left), quality-adjusted Life Years (QALYs, centre) and costs (right) of emergency surgery versus non-emergency surgery from the Local Instrumental Variables (LIV) approach for acute appendicitis (panel A) and acute gallstone disease (B)

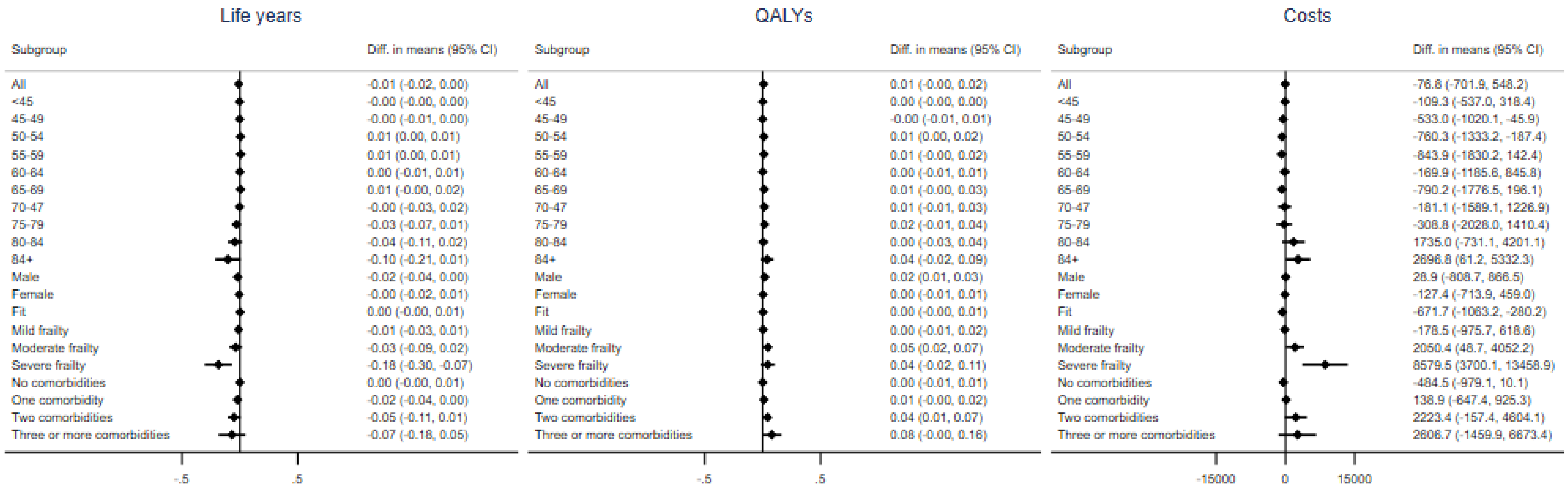
Figure 1. (A) Acute appendicitis



*Values to the left (right) of the 0 axis indicate that NES (ES) leads to fewer life years/QALYs (left and centre panel) or reduced costs (right panel) for the subgroup.

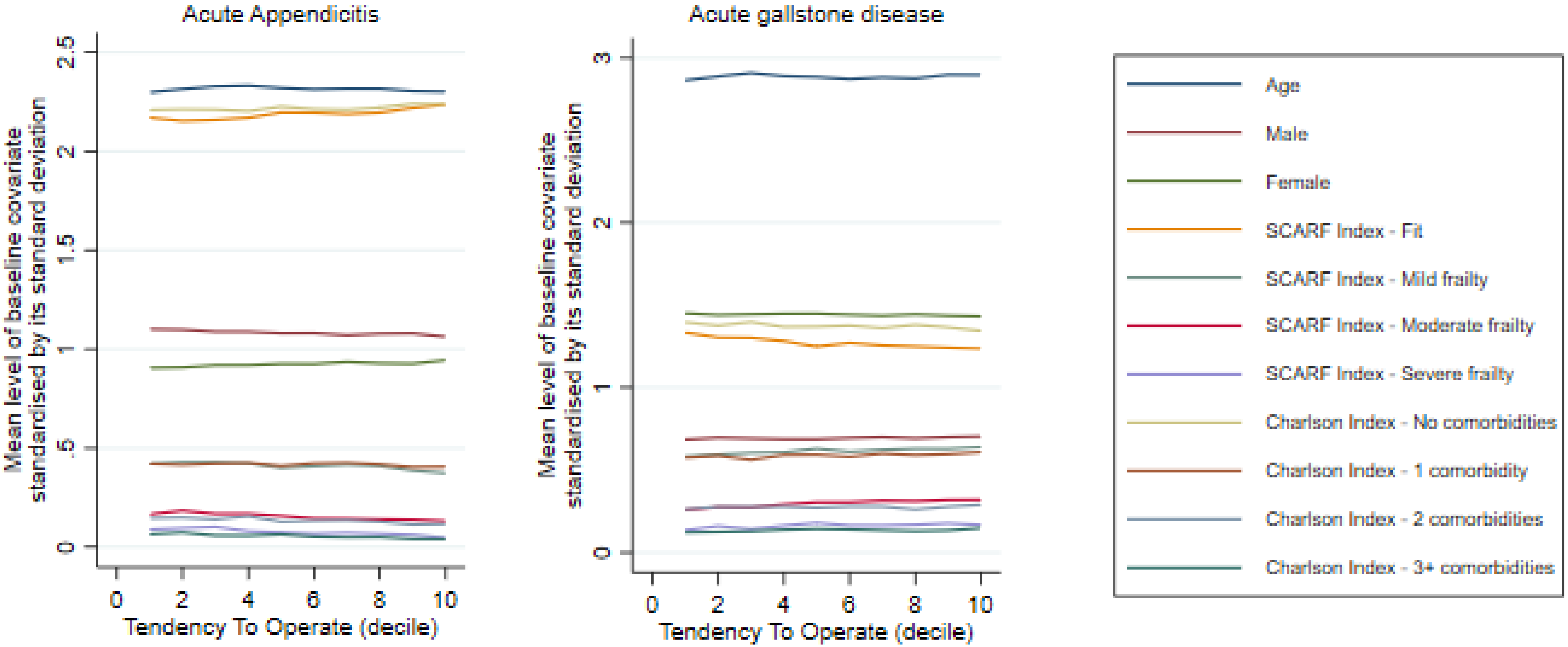
Appendix B.4. Forest plots of estimated incremental life years (left), quality-adjusted Life Years (QALYs, centre) and costs (right) of emergency surgery versus non-emergency surgery from the Local Instrumental Variables (LIV) approach for acute appendicitis (panel A) and acute gallstone disease (B)

Figure 1. (b) Acute gallstone disease



*Values to the left (right) of the 0 axis indicate that NES (ES) leads to fewer life years/QALYs (left and centre panel) or reduced costs (right panel) for the subgroup.

Appendix B.5. Mean level of rescaled baseline covariates according to the level of the instrumental variable



*SCARF: Secondary Care Administrative Records Frailty.

Appendix C. Chapter 4

Appendix C.1. Costing methodology

Overview

The main cost items for emergency surgery (ES) and non-emergency surgery (NES) strategies were the costs related to the length of stay in hospital, including the stay on general wards (bed-day costs) and on intensive care units (ICU). These costs covered both the index admission and any readmission (emergency or planned) up to one year, and used patient-level resource use data from HES, that linked information across all qualifying hospital admissions. The costs considered also included diagnostic and operative procedures. A micro-costing approach was adopted to calculate the costs of operative procedures in the index admission and readmissions.

Costing operative procedures

The study designated more common operative procedures as potential drivers of the incremental cost of ES versus NES, and therefore costed each of these procedures separately. For each of the comparison groups, the definition of 'more common' was a procedure with a prevalence that exceeded 1% in the index admission. This conservative definition of 'more common' was taken to reduce the risk of excluding important cost differences between the comparison groups. The 1% rule was deemed appropriate for also identifying 'more common' procedures in readmissions, which were costed using the same methods and assumptions used for procedures performed in the index admission. If multiple operative procedures met the criteria for 'more common', only one was costed. For each comparison group, we first considered procedures that would potentially qualify as ES (see Appendix C.6.D), and then considered operative procedures that did not meet the ES criteria (e.g. cholecystectomy) and then other non-operative procedures (e.g. catheterisation of bladder). For those admissions with procedures that did not meet the threshold for a common procedure because they were 'low-volume' (<1% for the comparison group in the initial admission) we did not include specific additional costs for the procedure per se, and only included the costs associated with bed-days and diagnostic tests[i].

[i] The same 1% rule was used for identifying the most common diagnostic procedures that were costed.

Appendix C.13 lists the most common operative procedures for each of the comparison groups for the three conditions.

To calculate the costs of the 'common' operative procedures, the expected durations of the operations, the number and grade of staff involved were informed by the literature and expert opinion (see Appendix C.9). The use of disposables (e.g. reload staplers), equipment (e.g. imaging systems), surgical instruments (e.g. laparoscopic sets), and overheads were informed by expert opinion (see Appendix C.9).

Applying unit costs

Each resource use item was valued using appropriate unit costs from recommended national sources (see Appendices C.7 y C.8). Direct personnel costs were calculated as the costs per hour of employing each grade of staff. The costs of overheads included costs of drugs, direct Central Sterile Supply Department (CSSD), as well as allocated costs (rent, property and equipment maintenance and cleaning costs, among others) associated with the provision of the procedure (ISD Scotland, 2019). Purchase prices of disposables, instruments and equipment for each procedure were retrieved from different sources, including the finance department of an NHS Trust hospital. The assignment of unit costs for each item took account of the expected number of times the item would be used over the lifetime, recognising any additional costs of a sterilisation process required to enable reuse (Ismail et al., 2015). All unit costs were inflated to 2019/20 prices (£ GBP) using UK's GDP deflator published by HM Treasury (HM Treasury Department, 2020).

To assess the sensitivity of the results to assumptions made about unit costs, the sensitivity analyses considered alternative scenarios. Specifically, the inclusion of the full unit costs of operative procedures risks double counting of those items (e.g., some consumables) that may be included within the overall costs per bed-day. Conversely, the exclusion of the 'less common' operative procedures for both comparison groups may have led to an underestimate of the absolute levels of costs for both groups. To investigate whether either standpoint would be likely to lead to a large inaccuracy in the estimation of incremental cost, the scenarios considered increases and decreases of by 10% (see Appendix C.14).

Appendix C.2. Calculation of QALYs

The cost-effectiveness analysis (CEA) was designed to report QALYs up to one year (base case) by combining individual-level survival data with appropriate health-related quality of life (HRQoL) estimates for acute appendicitis, diverticulitis and abdominal wall hernia, for ES and NES strategies.

For the base case analysis, HRQoL values were required at 'baseline', the time of the emergency admission, and at one-year follow-up. For survivors at one-year, it was assumed that patients' HRQoL was reduced for the duration of the initial emergency hospital admission, and then following hospital discharge that the patient's HRQoL level recovered immediately to the average HRQoL level reported in the literature at the one-year follow-up. For patients who had an emergency readmission recorded within the HES data, during the one-year follow-up, it was assumed that HRQoL at readmission reverted to the same level as that following the initial (index) emergency admission. It was also assumed that following hospital discharge the HRQoL levels reverted to those at one-year follow-up (see Appendix C.18). The assumption that HRQoL reverted to follow-up levels immediately after hospital discharge, was challenged in sensitivity analysis in which QALYs were instead calculated using linear interpolation between the index emergency admission and one-year follow-up (see Appendix C.18). For patients who died prior to one-year, a HRQoL score of zero was applied.

The approach to estimating QALYs therefore assumed that events that do not lead to emergency readmissions (e.g. planned surgery for recurrence), have minimal impact on the patient's HRQoL, as suggested previously e.g. for hernia repairs in the elective setting in McCormack et al., (2005), and Sharma et al., (2015). It was also assumed that there is no differential effect on QALYs between the comparison groups, beyond the effect on one-year mortality, or the rate or duration of emergency readmissions, both of which were derived from the individual-level HES data. This was motivated by the limited availability of studies comparing HRQoL of ES to NES alternatives in the emergency setting. The QALY calculation recognised differences in HRQoL according to age and gender, by adjusting the general HRQoL values from the literature according to recommended age-gender weights derived from the general population (see Appendix C.11) (Ara et al., 2017; Ara and Brazier, 2010).

Appendix C.3. Search for appropriate HRQoL scores and adjustment

The approach to estimating QALYs required that appropriate HRQoL values were identified from a literature review. We undertook separate search strategies for each condition in MEDLINE (see Appendix C.10). The criteria used to select the most appropriate source of HRQoL, recognised the specific requirements of the ESORT study (ESORT Study Group, 2020), and were prioritised according to:

1. The study considered at least one intervention regarded as ES by the clinical panel.
2. The intervention was performed in the emergency (non-elective) setting.
3. The study evaluated HRQoL using the tool recommended by NICE in their methodological guidance, the EuroQoL 5-dimension (EQ-5D) instrument in its three-level (3L) version (NICE, 2013).
4. The study evaluated HRQoL at baseline (i.e. pre-operatively), and at one-year from baseline.
5. The study was conducted in the UK, or in a country with similar demographics and healthcare system.
6. The study was conducted no earlier than ten years before the start date of the ESORT study (i.e. 2010).

Most of the studies that met these criteria compared different forms of ES, rather than ES versus NES strategies. We therefore applied the same HRQoL scores at baseline and one year to both comparison groups as outlined in Appendix C.2, in keeping with the assumption noted above, that any differences in HRQoL between the comparison groups would be captured by differences in one-year mortality, and the rate and duration of emergency readmissions.

Appendix C.4. A Local instrumental variable (LIV) approach

We consider the Neyman-Rubin potential outcomes framework (Neyman, 1990; Rubin, 1974), where $Y_1 = \mu_1(X_O, X_U, \vartheta)$ and $Y_0 = \mu_0(X_O, X_U, \vartheta)$ are the potential outcomes under treatments 1 (ES) and 0 (NES) and $\Delta = Y_1 - Y_0$ is the individual treatment effect, X_O are observed characteristics (e.g. patient's measured frailty), X_U are unmeasured confounders (e.g. patient's physiology) and ϑ captures any remaining unobserved random variation. The model for treatment assignment can be defined as $D^* = \mu_D(Z, X_O) - U_D$ and $D = 1\ if\ D^* \geq 0$, where Z is a vector of instruments and U_D captures X_U and any other unobserved variable that influences treatment selection. Here, the decision to assign the patient to ES (D=1) depends on their observed and unobserved characteristics and the tendency of their hospital to operate (i.e. the instrument). Following Heckman and Vytlacil (1999, 2005) and without loss of generality, this model can be re-written in terms of probabilities as $D^* = P(Z, X_O) - V$, where $P(Z, X_O)$ is the propensity for treatment, V reflects the degree to which unobserved variables discourage treatment and V is uniformly distributed between 0 and 1.

Note that under this model each complier (patient whose treatment status was altered by shifts in the level of IV) has some level of Z at which they would have only just been assigned to ES. For any value of Z below that "threshold", the patient will remain in the comparator group. At this level of Z they would be in equipoise. These hypothetical patients in equipoise are referred to as marginal patients, since a marginal change in the IV is sufficient to alter their treatment assignment. Since the IV is assumed not to otherwise influence treatment, the change in outcomes attributable to this marginal change in the IV can be attributed to the change in treatment, thus we can identify the marginal treatment effect for these marginal patients. The Marginal Treatment Effect (MTE) can be defined as,

$$\Delta^{MTE}(x_O, v) = E(\Delta | X_O = x_O, V = v)$$

The MTE is the most nuanced treatment effect parameter. Under regular IV assumptions, the Local instrumental variable (LIV) estimator can be used to estimate a series of MTEs (Bjorklund and Moffitt, 1987; Heckman, 1997; Heckman and Vytlacil, 1999, 2005),

$$\Delta^{MTE}(x_O, p) = \frac{\partial E(Y_1 - Y_0 | X_O = x_O, P(z, x_O) = p)}{\partial p}$$

To estimate the MTEs, we (i) estimate the propensity score for ES for each individual using a probit model, including the measured confounders and the instrument (similar to the first stage in 2SRI, or 2SLS but using probit in place of linear regression), (ii) store the estimated propensity scores and make sure that there exists coverage for both treatment arms across all values from 0 to 1 (rounded to 0.01; or else drop values), (iii) estimate an outcome model (Generalised Linear Models (GLMs) here) on the covariates, propensity score and interactions of these using appropriate methods, (iv) take the derivative of the estimated outcome equation with respect to the estimated propensity score to obtain the MTE estimate. Then, MTEs can be aggregated into meaningful parameters of treatment effects such as the ATE or CATEs.

Basu showed that MTEs can also be used to derive person-centered treatment (PeT) effects (Basu, 2014, 2015). Note that the treatment assignment status provides some information on V and $P(Z, X_O)$ for each patient. For instance, a patient with low frailty who nonetheless is assigned to ES, is likely to have unmeasured characteristics encouraging ES (disease severity), that is the observed treatment assignment is informative about the range of unobserved confounders (and hence values of $V)$ that are plausible for that patient.

For patients in the treatment group (D=1), the propensity to choose treatment based on X and Z must outweigh the propensity to choose the comparator strategy based on U_D, i.e. $P(z, x_O) > v$. For patients in the comparator strategy (D=0), the opposite is true. Hence,

$$\Delta^{PeT}(x_O, p, D) = E(Y_1 - Y_0 | X_O = x_O, P(z, x_O) > v) \text{ for individuals with D=1}$$

$$\Delta^{PeT}(x_O, p, D) = E(Y_1 - Y_0 | X_O = x_O, P(z, x_O) < v) \text{ for individuals with D=0}$$

The PeT effect averages MTEs with the same level of X and Z over those values of unobserved variables that are compatible with that patient's treatment assignment. All the treatment effect parameters, including conditional average treatment effects (CATEs), can be derived by taking averages of PeT effects. This can be accomplished using the 'petiv' command in Stata.(Basu, 2015) In short, we evaluate the MTE at different values of v, retaining only those that are consistent with the observed treatment decision given that patients observed characteristics and the level of the hospital's tendency to operate (TTO, i.e. the IV) for their hospital, and then average these MTEs to obtain the PeT effect. The PeT effects can then be aggregated for the population of interest.

Implementation for the CEA

This LIV approach was implemented as follows: first, each patient's propensity for ES was estimated according to their observed characteristics and the TTO using a probit model. Second, for each outcome (costs, QALYs), an appropriate GLM, determined by reference to the root mean squared error, was estimated relating the observed outcome to the individuals' observed characteristics, and their propensity for ES, along with interactions between them. Next, the MTEs were obtained by considering the impact on outcomes of a marginal change in the propensity for ES. Third, numerical integration was used to obtain individual level treatment effect estimates recognising their actual treatment assignment as described in Basu (2015). After obtaining the effect estimates for Costs and QALYs, these were used to calculate the effect on Net Monetary Benefit (NMB), i.e. the incremental net monetary benefit (INB).

To obtain standard errors and confidence intervals, the steps above were bootstrapped 300 times (200 times for sensitivity analyses due to computational complexity), with all outcome models estimated within the same bootstrap to account for correlation between the cost and QALY endpoints. (For further details on the estimation steps, see Basu (2015)).

Appendix C.5. Accounting for hospital quality

We derived proxy measures for the quality of acute care in managing emergency admissions. These proxy measure of quality of care, were defined by the rates of all-cause mortality and emergency readmissions up to 90 days for each hospital (base case). This information was reported for each condition for the 2009-2010 financial year, to provide baseline, time-invariant proxies for care quality in each hospital, and for the one year preceding each qualifying emergency hospital admission, to provide time-varying proxies for care quality. This allowed the study to adjust for time-constant differences in quality across hospitals, and those that differed over time. While an alternative approach would be to include hospital level fixed effects, these would only control for time invariant unobserved confounders, and would also remove much of the variation in TTO by hospital, thus weakening the IV substantially.

In sensitivity analyses, we consider 'external' measures of 'quality of acute care' by using hospital performance measures from the National Emergency Laparotomy Audit (NELA)(NELA, 2016, 2017, 2018). Since data were not available from NELA for all years of the study, and definitions changed over time, we constructed an average (weighted by volume) using data from 2016, 2017 and 2018 for the following seven indicators of quality of peri-operative management for emergency laparotomy patients which we anticipate would capture the influence of any potential time invariant observed confounders associated with hospital quality:

1. Adjusted mortality rate
2. Proportion of patients in whom a risk assessment was documented preoperatively
3. Proportion of patients arriving in theatre within a time appropriate for the urgency of surgery
4. Proportion of patients with a calculated preoperative risk of death >5% for whom a consultant surgeon and anaesthetist were present in theatre
5. Admission to critical care when risk of death ≥5%
6. Unplanned returns to theatre
7. Unplanned returns to critical care

These variables were anticipated to control for a range of potential hospital-level unobserved confounders.

Appendix C.6. Definitions of populations (panel A and B) and interventions (C and D) for acute appendicitis, diverticular disease, and abdominal wall hernia

(A): List of International Classification of Diseases (ICD)-10 codes considered for inclusion criteria

Acute appendicitis (N=268,144)	**Diverticular disease (N=138,869)**	**Abdominal wall hernia (N=106,432)**
K35: Acute appendicitis	K57.0: Diverticular disease of small intestine with perforation and abscess	K40.0: Bilateral inguinal hernia, with obstruction, without gangrene
K35.2: Acute appendicitis with generalised peritonitis	K57.1: Diverticular disease of small intestine without perforation or abscess	K40.1: Bilateral inguinal hernia, with gangrene
K35.3: Acute appendicitis with localized peritonitis	K57.2: Diverticular disease of large intestine with perforation and abscess	K40.2: Bilateral inguinal hernia, without obstruction or gangrene
K35.8: Acute appendicitis, other and unspecified	K57.3: Diverticular disease of large intestine without perforation or abscess	K40.3: Unilateral or unspecified inguinal hernia, with obstruction, without gangrene
K37: Unspecified appendicitis	K57.4: Diverticular disease of both small and large intestine with perforation and abscess	K40.4: Unilateral or unspecified inguinal hernia, with gangrene
	K57.5: Diverticular disease of both small and large intestine without perforation or abscess	K40.9: Unilateral or unspecified inguinal hernia, without obstruction or gangrene
	K57.8: Diverticular disease of intestine, part unspecified, with perforation and abscess	K41.0: Bilateral femoral hernia, with obstruction, without gangrene
	K57.9: Diverticular disease of intestine, part unspecified, without perforation or abscess	K41.1: Bilateral femoral hernia, with gangrene
		K41.2: Bilateral femoral hernia, without obstruction or gangrene
		K41.3: Unilateral or unspecified femoral hernia, with obstruction, without gangrene

Acute appendicitis (N=268,144)	Diverticular disease (N=138,869)	Abdominal wall hernia (N=106,432)
		K41.4: Unilateral or unspecified femoral hernia, with gangrene K41.9: Unilateral or unspecified femoral hernia, without obstruction or gangrene K42.0: Umbilical hernia with obstruction, without gangrene K42.1: Umbilical hernia with gangrene K42.9: Umbilical hernia without obstruction or gangrene K43.0: Incisional hernia with obstruction, without gangrene K43.1: Incisional hernia with gangrene K43.2: Incisional hernia without obstruction or gangrene K43.3: Parastomal hernia with obstruction, without gangrene K43.4: Parastomal hernia with gangrene K43.5: Parastomal hernia without obstruction or gangrene K43.6: Other and unspecified ventral hernia with obstruction, without gangrene K43.7: Other and unspecified ventral hernia with gangrene K43.9: Other and unspecified ventral hernia without obstruction or gangrene

Appendix C.6. Definitions of populations (panel A and B) and interventions (C and D) for acute appendicitis, diverticular disease, and abdominal wall hernia

(B): List of exclusion criteria

Acute appendicitis (N=268,144)	Diverticular disease (N=138,869)	Abdominal wall hernia (N=106,432)
Pregnancy **Appendiceal cancer**	None	Pregnancy Ischaemia Cancer

(C): Definition of 'emergency surgery' and time window

	Acute appendicitis (N=268,144)	Diverticular disease (N=138,869)	Abdominal wall hernia (N=106,432)
Procedures defined as 'emergency surgery'	See Panel (D)	See Panel (D)	See Panel (D)
Common procedures excluded from definition of 'emergency surgery'	*Unspecified other excision of appendix	Image controlled percutaneous drainage	None
Threshold for a procedure in the index admission to be 'emergency surgery'	7 days	Any time	3 days
Threshold for a procedure in a readmission to be 'emergency surgery'	7 days	14 days	3 days

*Further OPCS Classification of Interventions and Procedures (OPCS-4) codes were added to the list of ES procedures after the clinical panel exercise. For appendicitis (following review of coding use by hospital): H029 Unspecified other excision of appendix. For abdominal wall hernia (following inclusion of umbilical hernia as a diagnosis and for consistency with other hernia types): T241 Repair of umbilical hernia using insert of natural material, T248 Other specified primary repair of umbilical hernia, T971 Repair of recurrent umbilical hernia using insert of natural material, T973 Repair of recurrent umbilical hernia using sutures, T978 Other specified repair of recurrent umbilical hernia, T979 Unspecified repair of recurrent umbilical hernia. See Panel (D) for full list of OPCS codes defined as emergency surgery.

Appendix C.6. Definitions of populations (panel A and B) and interventions (C and D) for acute appendicitis, diverticular disease, and abdominal wall hernia

(D): Full list of OPCS codes defined as emergency surgery

Acute appendicitis (N=268,144)	**Diverticular disease (N=138,869)**	**Abdominal wall hernia (N=106,432)**
H011: Emergency excision of abnormal appendix and drainage H012: Emergency excision of abnormal appendix NEC H018: Other specified emergency excision of appendix H019: Unspecified emergency excision of appendix H029: Unspecified other excision of appendix H031: Drainage of abscess of appendix H032: Drainage of appendix H071: Right hemicolectomy and end to end anastomosis of ileum to colon H072: Right hemicolectomy and side to side anastomosis of ileum to transverse colon H073: Right hemicolectomy and anastomosis NEC H074: Right hemicolectomy and ileostomy HFQ T342: Open drainage of pelvic abscess T343: Open drainage of abdominal abscess H013: Emergency excision of normal appendix T463: Irrigation of peritoneal cavity H062: Extended right hemicolectomy and anastomosis of ileum to colon	H091: Left hemicolectomy and end to end anastomosis of colon to rectum H092: Left hemicolectomy and end to end anastomosis of colon to colon H093: Left hemicolectomy and anastomosis H094: Left hemicolectomy and ileostomy H095: Left hemicolectomy and exteriorisation of bowel NEC H101: Sigmoid colectomy and end to end anastomosis of ileum to rectum (0.03%) H102: Sigmoid colectomy and anastomosis of colon to rectum H103: Sigmoid colectomy and anastomosis H104: Sigmoid colectomy and ileostomy H105: Sigmoid colectomy and exteriorisation of bowel H113: Colectomy and anastomosis NEC (0.01%) H114: Colectomy and ileostomy H115: Colectomy and exteriorisation of bowel H152: End colostomy H158: Other specified other exteriorisation of colon H333: Anterior resection of rectum and anastomosis of colon to rectum using staples	T201: Primary repair of inguinal hernia using insert of natural material T202: Primary repair of inguinal hernia using insert of prosthetic material T203: Primary repair of inguinal hernia using sutures T204: Primary repair of inguinal hernia and reduction of sliding hernia T208: Other specified primary repair of inguinal hernia T209: Unspecified primary repair of inguinal hernia T211: Repair of recurrent inguinal hernia using insert of natural material T212: Repair of recurrent inguinal hernia using insert of prosthetic material T213: Repair of recurrent inguinal hernia using sutures T218: Other specified repair of recurrent inguinal hernia T219: Unspecified repair of recurrent inguinal hernia T221: Primary repair of femoral hernia using insert of natural material T222: Primary repair of femoral hernia using insert of prosthetic material T223: Primary repair of femoral hernia using sutures

Acute appendicitis (N=268,144)	**Diverticular disease (N=138,869)**	**Abdominal wall hernia (N=106,432)**
H078: Other specified other excision of right hemicolon T468: Other specified other drainage of peritoneal cavity	H334: Anterior resection of rectum and anastomosis NEC H335: Rectosigmoidectomy and closure of rectal stump and exteriorisation of bowel H336: Anterior resection of rectum and exteriorisation of T342: Open drainage of pelvic abscess (0.09%) 33	T228: Other specified primary repair of femoral hernia T229: Unspecified primary repair of femoral hernia T231: Repair of recurrent femoral hernia using insert of natural material T232: Repair of recurrent femoral hernia using insert of prosthetic material T233: Repair of recurrent femoral hernia using sutures T239: Unspecified repair of recurrent femoral hernia T241: Repair of umbilical hernia using insert of natural material T242: Repair of umbilical hernia using insert of prosthetic material T243: Repair of umbilical hernia using sutures T248: Other specified primary repair of umbilical hernia T249: Unspecified primary repair of umbilical hernia T271: Repair of ventral hernia using insert of natural material T272: Repair of ventral hernia using insert of prosthetic material T273: Repair of ventral hernia using sutures T278: Other specified repair of other hernia of abdominal wall T279: Unspecified repair of other hernia of abdominal wall

Acute appendicitis (N=268,144)	**Diverticular disease (N=138,869)**	**Abdominal wall hernia (N=106,432)**
		T288: Other specified other repair of anterior abdominal wall G762: Open relief of strangulation of ileum G763: Open relief of obstruction of ileum NEC H176: Open relief of obstruction of colon NEC T251: Primary repair of incisional hernia using insert of natural material T971: Repair of recurrent umbilical hernia using insert of natural material T972: Repair of recurrent umbilical hernia using insert of prosthetic material T973: Repair of recurrent umbilical hernia using sutures T978: Other specified repair of recurrent umbilical hernia T979: Unspecified repair of recurrent umbilical hernia T981: Repair of recurrent ventral hernia using insert of natural material T982: Repair of recurrent ventral hernia using insert of prosthetic material T983: Repair of recurrent ventral hernia using sutures T989: Unspecified repair of recurrent other hernia of abdominal wall T252: Primary repair of incisional hernia using insert of prosthetic material T253: Primary repair of incisional hernia using sutures

Acute appendicitis (N=268,144)	**Diverticular disease (N=138,869)**	**Abdominal wall hernia (N=106,432)**
		T258: Other specified primary repair of incisional hernia T259: Unspecified primary repair of incisional hernia 44 T261: Repair of recurrent incisional hernia using insert of natural material T262: Repair of recurrent incisional hernia using insert of prosthetic material T263: Repair of recurrent incisional hernia using sutures T268: Other specified repair of recurrent incisional hernia T269: Unspecified repair of recurrent incisional hernia T318: Other specified other operations on anterior abdominal wall

Appendix C.7. Unit costs (£GBP 2019/20) for potential cost drivers

Item	Unit	Unit cost (£GBP)	Source, definitions and assumptions
		Inpatient stay	
General ward	Day	347	NHS Reference costs 2017/18. Weighted average of FD05A and FD05B (NEL_XS) (NHS Improvement, 2018).
		ICU ward	
Level 2 ICU	Day	1,188	NHS Reference costs 2017/18. XC06Z: 1 organ supported (adult critical care) (NHS Improvement, 2018).
Level 3 ICU	Day	1,886	NHS Reference costs 2017/18. Weighted average of XC01Z-XC05Z. 2 to 6+ organs supported (adult critical care) .
		Diagnostic procedures	
More common diagnostic procedures for acute appendicitis			
Computed tomography	Procedure	83	NHS Reference costs 2017/18. RD20A: Computerised Tomography Scan of One Area, without Contrast, 19 years and over (IMAG) (NHS Improvement, 2018).
Unspecified diagnostic endoscopic examination of colon	Procedure	206	NHS Reference costs 2017/18. FE31Z: Diagnostic Colonoscopy with Biopsy, 19 years and over (NES). Mean bed-day costs of general ward subtracted to avoid double-counting (NHS Improvement, 2018).
Fibreoptic endoscopic examination of upper gastrointestinal tract and biopsy of lesion of upper gastrointestinal tract	Procedure	197	NHS Reference costs 2017/18. FE21Z: Diagnostic Endoscopic Upper Gastrointestinal Tract Procedures with Biopsy, 19 years and over (NES). Mean bed-day costs of general ward subtracted to avoid double-counting (NHS Improvement, 2018).
Diagnostic fibreoptic endoscopic	Procedure	277	NHS Reference costs 2017/18. FE31Z: Diagnostic Colonoscopy with

Item	Unit	Unit cost (£GBP)	Source, definitions and assumptions
examination of colon and biopsy of lesion of colon			Biopsy, 19 years and over (NES). Mean bed-day costs of general ward subtracted to avoid double-counting (NHS Improvement, 2018).
Computed tomography of head	Procedure	83	NHS Reference costs 2017/18. RD20: Computerised Tomography Scan of One Area, without Contrast, 19 years and over (IMAG) (NHS Improvement, 2018).
More common diagnostic procedures for diverticular disease			
Computed tomography	Procedure	83	NHS Reference costs 2017/18. RD20: Computerised Tomography Scan of One Area, without Contrast, 19 years and over (IMAG) (NHS Improvement, 2018).
Unspecified diagnostic endoscopic examination of lower bowel using fibreoptic sigmoidoscope	Procedure	143	NHS Reference costs 2017/18. FE35Z: Diagnostic Flexible Sigmoidoscopy, 19 years and over (NES). Mean bed-day costs of general ward subtracted to avoid double-counting (NHS Improvement, 2018).
Unspecified diagnostic endoscopic examination of colon	Procedure	206	NHS Reference costs 2017/18. FE32Z: Diagnostic Colonoscopy, 19 years and over (NES). Mean bed-day costs of general ward subtracted to avoid double-counting (NHS Improvement, 2018).
Unspecified diagnostic fibreoptic endoscopic examination of upper gastrointestinal tract	Procedure	277	NHS Reference costs 2017/18. FE31Z: Diagnostic Colonoscopy with Biopsy, 19 years and over (NES). Mean bed-day costs of general ward subtracted to avoid double-counting (NHS Improvement, 2018).
Diagnostic endoscopic examination of lower bowel and biopsy of lesion of lower bowel using fibreoptic sigmoidoscope	Procedure	205	NHS Reference costs 2017/18. FE34Z: Diagnostic Flexible Sigmoidoscopy with Biopsy, 19 years and over (NES). Mean bed-day costs of general ward (see above) subtracted to avoid double-counting (NHS Improvement, 2018).
More common diagnostic procedures for abdominal wall hernia			

Item	Unit	Unit cost (£GBP)	Source, definitions and assumptions
Computed tomography	Procedure	83	NHS Reference costs 2017/18. RD20A: Computerised Tomography Scan of One Area, without Contrast, 19 years and over (IMAG) (NHS Improvement, 2018).
Transthoracic echocardiography	Procedure	101	NHS Reference costs 2017/18. RD51C: Simple Echocardiogram, 5 years and under (IMAG) (NHS Improvement, 2018).
Computed tomography of abdomen	Procedure	83	NHS Reference costs 2017/18. RD20: Computerised Tomography Scan of One Area, without Contrast, 19 years and over (IMAG) (NHS Improvement, 2018).
Computed tomography of head	Procedure	83	NHS Reference costs 2017/18. RD20A: Computerised Tomography Scan of One Area, without Contrast, 19 years and over (IMAG) (NHS Improvement, 2018).
Diagnostic endoscopic examination of peritoneum	Procedure	404	NHS Reference costs 2017/18. FE31Z: Diagnostic Colonoscopy with Biopsy, 19 years and over (NES). Mean bed-day costs of general ward subtracted to avoid double-counting (NHS Improvement, 2018).
		Operative procedures	
Staff input			
Consultant surgeon	Minute	1.8	2019 Unit costs of Health and Social Care (PSSRU). Section 14. Cost per working hour: consultant: surgical (Curtis and Burns, 2019).
Anaesthesiologist	Minute	1.8	2019 Unit costs of Health and Social Care (PSSRU). Section 14. Cost per working hour: consultant: medical (Curtis and Burns, 2019)
Consultant radiologist	Minute	1.8	2019 Unit costs of Health and Social Care (PSSRU). Section 14. Cost per working hour: consultant: medical (Curtis and Burns, 2019).
Registrar – surgery	Minute	0.8	2019 Unit costs of Health and Social Care (PSSRU). Section 14. Cost per

Item	**Unit**	**Unit cost (£GBP)**	**Source, definitions and assumptions**
			working hour: registrar (Curtis and Burns, 2019).
Registrar – anaesthesiology	Minute	0.8	2019 Unit costs of Health and Social Care (PSSRU). Section 14. Cost per working hour: registrar (Curtis and Burns, 2019).
Registrar – radiology	Minute	0.8	2019 Unit costs of Health and Social Care (PSSRU). Section 14. Cost per working hour: registrar (Curtis and Burns, 2019).
Nurse – Band 5	Minute	0.6	2019 Unit costs of Health and Social Care (PSSRU). Section 13. Cost per working hour: band 5 – hospital-based nurse (Curtis and Burns, 2019).
Nurse – Band 6	Minute	0.8	2019 Unit costs of Health and Social Care (PSSRU). Section 13. Cost per working. hour: band 6 – hospital-based nurse (Curtis and Burns, 2019).
Operating department practitioner	Minute	0.8	2019 Unit costs of Health and Social Care (PSSRU). Section 13. Assumed same cost as cost per working hour of band-6 hospital-based nurse (Chapter 13) (Curtis and Burns, 2019).
Overhead costs			
Operating room	Minute	5.4	Includes direct drug and CSSD costs as well allocated costs (other staff; property and equipment maintenance; domestics and cleaning; heat, light and power; rent and rates; purchases of furniture, fittings and equipment (non-capital charge) and others). Weighted average of 43 hospitals in Scotland (ISD Scotland., 2019).
Reusable instruments and equipment			
Laparoscopic colorectal set	Procedure	39.2	Manufacturer. See Table S3 for full list of components. Total purchase cost is £3,112. Number of uses is 2,750. Final cost includes sterilisation cost following at £0.8 cost per instrument used (Ismail et al., 2015).
Main laparoscopic set	Procedure	36.8	Manufacturer. See Table S3 for full list of components. Total purchase cost is £2,511. Assumed number of

Item	Unit	Unit cost (£GBP)	Source, definitions and assumptions
			uses is 2,750. Final cost includes sterilisation cost following at £0.8 cost per instrument used (Ismail et al., 2015).
Major general set	Procedure	39.2	Manufacturer. See Table S3 for full list of components. Total purchase cost is £2,744. Assumed number of uses is 2,750. Final cost includes sterilisation cost following at £0.8 cost per instrument used (Ismail et al., 2015).
Minor general set	Procedure	32.8	Manufacturer. See Table S3 for full list of components. Total purchase cost is £1,417. Assumed number of uses is 2,750. Final cost includes sterilisation cost following at £0.8 cost per instrument used (Ismail et al., 2015).
Endoscopic polypectomy set	Procedure	16.2	Manufacturer. Includes endoscopic forceps, snare and endoscopic clips. Final cost includes sterilisation cost following at £0.8 cost per instrument used. (Ismail et al., 2015) Unit cost calculated assuming number of uses is 4400 (except for snare and clips which are assumed to be disposable).
Telescope and stack	Procedure	15.2	Manufacturer. Includes stack, scope (Precision ideal eyes 10mm 30°, HD autoclavable Laparoscope 33cm), tray and cable (fibreoptic cable 5.0mm x 10 ft. (3.05m)). Purchase cost of stack and stack are £68,760 and £2,334, respectively. Unit cost calculated assuming number of uses is 4400.
Ultrasound system	Procedure	1.5	Manufacturer. Purchase cost of ultrasound system is £7,132. Unit cost calculated assuming number of uses is 4400.
Disposables			
Laparoscopic linear stapler	Procedure	262	Manufacturer. Linear Cutter 75mm. 1 is assumed to be used per procedure.

Item	Unit	Unit cost (£GBP)	Source, definitions and assumptions
Stapler reload	Procedure	36.5	Manufacturer. Reload linear cutter, blue, 75mm. Purchase cost of £465.61 per box of 12. 1 is assumed to be used per procedure.
Endoloop ligature	Procedure	56.9	Manufacturer. Endoloop Ethicon. 3 are assumed to be used per procedure (Clement et al., 2020).
Biosynthetic mesh	Procedure	61.1	Manufacturer. Sutumed Polipropilene Non-absorbable Hernia Mesh 12" X 12". 1 is assumed to be used per procedure.
Abdominal drain set	Procedure	18.2	Manufacturer. Set includes 1000mL drainage bag, catheter valve cap, slide clamp, tape strips and wipe. Purchase cost £36.5 per box of 2.
Foyle catheterisation kit	Procedure	9.4	Manufacturer. Catheterisation Set 16fr Foley and extras. Includes a 16fr Foley catheter, a 500ml leg-bag, 2000ml bedside drainage bag, sterile syringe and lube.

CSSD: Central sterile services department, ICU: intensive care unit, NHS: National Health Service.

Appendix C.8. Full list of components of surgical sets considered in cost analysis

Laparoscopic colorectal set	Main laparoscopic set	Major general set	Minor general set
Aesculap dorsey forcep 4 parts	Anti-tamper tags	B p handle no 4	Artery forcep mosquito curved
Anti-tamper tags	B p handle no 3	B p handle no 5	B p handle no 3
B p handle no 3	B p handle no 4	Babcock tissue forcep 6 1/2"	B p handle no 4
B p handle no 4	Bottom tray	Babcock tissue forcep 8"	Babcock tissue forcep
Babcock tissue forcep long	Container	Balfour self-retaining retractor (see remarks)	Catspaw retractor
Babcock tissue forcep short	Container identification label	Deaver retractor, broad	Diathermy dissecting forcep mcindoe
Bottom tray	De-jardin stone forcep	Deaver retractor, narrow	Diathermy quiver
Container	Diathermy dissecting forcep mcindoe	Diathermy dissecting forcep mcindoe	Disposable blue tray wrap 120 x 150
Container identification label	Diathermy quiver long + black end cap	Diathermy quiver	Disposable green tray wrap 120 x 150
Diathermy dissecting forcep mcindoe	Dissecting forcep debakey 6"	Disposable green tray wrap 120 x 150	Dissecting forcep debakey 6"
Diathermy quiver	Dissecting forcep gillies toothed	Dissecting forcep debakey 6"	Dissecting forcep gillies toothed
Diathermy quiver long + black end cap	Dunhill artery forcep	Dissecting forcep debakey 8"	Dunhill artery forcep
Dissecting forcep debakey 6"	Eragon ratchet handle - do not assemble to forcep	Dissecting forcep debakey 9 1/2"	Heiss artery forcep
Dissecting forcep debakey 8"	Filter and retaining clip	Dissecting forcep gillies toothed	Lahey artery forceps
Dissecting forcep gillies toothed	Grasping forcep + ratchet with connector (pm 109)	Dissecting forcep non toothed 5"	Lanes dissecting forcep (1-2 teeth)
Doyen intestinal clamp curved	Hassan 10mm (2 parts+10mm clear seal)	Doyen intestinal clamp curved	Littlewoods tissue forcep
Dunhill artery forcep	Insulated hook with connector	Doyen intestinal clamp straight	Mayo pin holding next 2 items
Dyball retractor			Mayo pin holding next 3 items
Filter and retaining clip			Mayo pin holding next 4 items

Laparoscopic colorectal set	Main laparoscopic set	Major general set	Minor general set
Grasping forcep + ratchet with connector (pm 109)	Lanes dissecting forcep (1-2 teeth)	Dunhill artery forcep	Meyarding finger retractor
Hasson 12mm (3 parts) ea12nh send disassembled	Laparoscopic diathermy lead (8mm bovie)	Dyball retractor	Monopolar diathermy lead pin fitting
Heiss artery forcep	Littlewoods tissue forcep	Heiss artery forcep	Needle holder crilewood
Insulated hook with connector	Maryland forcep no ratchet with connector (pm 102)	Lahey artery forceps	Needle holder mayo hegar
Ireusable cannula 12mm	Mesh basket with lid	Lanes dissecting forcep (1-2 teeth)	Poirers/allis tissue forcep
Lanes dissecting forcep (1-2 teeth)	Modular monopolar forcep (johan) sn 8393.184 2 parts	Lang stevenson intestinal clamps	Retractor langenbeck medium
Laparoscopic diathermy lead (8mm bovie)	Monopolar diathermy lead pin fitting	Littlewoods tissue forcep	Retractor langenbeck small
Littlewoods tissue forcep	Myoma forcep + ratchet with connector (pm 117)	Massons needle holder	Retractor morris medium
Maryland f/cep no ratchet with connector (pm 102)	Needle holder crilewood	Mayo pin holding next 1 item	Retractor self-retaining travers
Massons needle holder	Needle holder mayo hegar	Mayo pin holding next 3 items	Retractor self-retaining west
Monopolar diathermy lead pin fitting	Pike mouth forcep + ratchet with connector (pm 107)	Mayo pin holding next 4 items	Scissor kilner curved
Needle holder mayo hegar	Retractor langenbeck medium	Mayo pin holding next 6 items	Scissor mayo curved
Nelson robert scissors	Retractor langenbeck small	Monopolar diathermy lead pin fitting	Scissor mayo straight 5 3/4"
Parker kerr intestinal clamp straight	Reusable cannula 10mm	Moynihan cholecystectomy clamp	Scissor mcindoe curved
Retractor langenbeck medium	Reusable cannula 12mm	Needle holder mayo hegar 7 1/4"	Sh/sh scissor
Retractor langenbeck small	Scissor mayo straight	Needle holder mayo hegar 8 1/2"	Soaker sheet to be placed under basket/tray
		Nelson robert scissors	Spencer wells artery forceps 7" curved
		Parker kerr intestinal clamp curved	Spencer wells artery forceps 8" straight
			Sponge holder rampley

Laparoscopic colorectal set	Main laparoscopic set	Major general set	Minor general set
Retractor morris medium Roberts artery forcep Scissor mayo straight Scissor mcindoe curved Sh/sh scissor Sponge holder rampley Threaded cannula 5mm (2 parts) Top tray Trayliner Trocar blunt tip 12mm Trocar pencil point 12mm Trocar pencil point 5mm Trocar sharp tip 5mm Waughs diathermy dissecting forcep	Scissor mcindoe curved Sh/sh scissor Spencer wells artery forceps 7" curved Sponge holder rampley Threaded cannula 5mm (2 parts) Top tray Towel clip small Trayliner Trocar blunt tip 10mm Trocar pencil point 12mm Trocar pencil point 5mm Trocar sharp tip 5mm Wash basket	Parker kerr intestinal clamp straight Retractor langenbeck medium Retractor morris large Roberts artery forcep Scissor mayo curved Scissor mayo straight 5 3/4" Scissor mcindoe curved Sh/sh scissor Soaker sheet to be placed under basket/tray Sponge holder rampley Styles tissue forcep Trayliner Wash basket Waughs diathermy dissecting forcep	T.o.e. dissecting forcep Trayliner Wash basket

Appendix C.9. Resource use categories for operative procedures in emergency surgery (ES) window

		Acute appendicitis		Diverticular disease		Abdominal wall hernia	
		ES (N=247,506)	NES (N=20,638)	ES (N=15,772)	NES (N=123,097)	ES (N=62,559)	NES (N=43,873)
Most common operative procedures in each arm in ES window	-	Emergency excision of abnormal appendix	Interval appendicectomy	Recto-sigmoidectomy and closure of rectal stump and exteriorisation of bowel	Fibreoptic endoscopic snare resection of lesion of colon	Primary repair of inguinal hernia using insert of prosthetic material	Unspecified urethral catheterisation of bladder
Time in theatre in minutes (source)	Literature /expert opinion	70 (Javanmard-Emamghissi et al., 2020)	70 (Javanmard-Emamghissi et al., 2020)	135 (Heah et al., 1995)*	25 (Teramoto et al., 2020)	60 (Wu et al., 2016)	15 (Wilson, 2016)
Staffing levels	Expert opinion	S1	S1	S1	S1	S1	S2
Instruments	Expert opinion	Main laparoscopic set	Main laparoscopic set	Major general set	Endoscopic polypectomy set	Minor general set	-
Equipment	Expert opinion	Laparoscope, cable and tray	Laparoscope, cable and tray	-	Laparoscope, cable and tray	-	-
Main disposables	Expert opinion	Three loops for closure of the appendiceal stump	Three loops for closure of the appendiceal stump	Laparoscopic linear stapler and reload	-	Biosynthetic mesh	Foyle catheterisation kit

The table includes exemplar data for most common operative procedures in ES window. Resource use for all other operative procedures was calculated considering the same categories. *If the procedure appeared with operative codes for loop colostomy, other specified other exteriorisation of colon, or unspecified other exteriorisation of colon, the duration was assumed to be 205 minutes. S1 considered: 1 consultant surgeon, 1 registrar surgeon, 2 band 5 nurses, 1 band 6 nurse, 1 operating department practitioner, 1 consultant anaesthetist, 1 registrar anaesthetist. S2 considered 1 band 5 nurse.

Appendix C.10. Search strategies for HRQoL data

(A): Acute appendicitis

```
Database: Ovid MEDLINE(R) ALL <1946 to August 19, 2021>
Search Strategy:
--------------------------------------------------------------------------------
1    *appendicitis/ (16067)
2    *appendectomy/ (6399)
3    appendic*.ti,ab. (33073)
4    appendec*.ti,ab. (10129)
5    emergency+surgery*.ti,ab. (9412)
6    emergency+appendectomy*.ti,ab. (153)
7    non-operative+manag*.mp. (1888)
8    conservative+manag*.mp. (16648)
9    antibiotic*.ti,ab. (360099)
10    antibiotic+adj+therapy.ti,ab. (0)
11    Anti-Bacterial+Agents/tu (135940)
12    Watchful+wait$.tu. (0)
13    delayed+surg$.ti,ab. (2186)
14    trial.ti,ab. (657219)
15    RCT.ti,ab. (24987)
16    randomi#ed+controlled+trial.pt. (541163)
17    controlled+clinical+trial.pt. (94345)
18    case+control+stud$.ti,ab. (113048)
19    cross-sectional+stud$.ti,ab. (197037)
20    cohort+stud$.ti,ab. (244943)
21    observational+stud$.ti,ab. (126477)
24    Economic+evaluation.ti,ab. (9983)
25    EuroQol-5+Dimension.ti,ab. (670)
26    "EQ-5D".ab. (9451)
27    or/1-2 (19106)
28    or/3-13 (496587)
29    and/27-28 (16860)
30    or/14-24 (1682168)
31    and/29-30 (1350)
32    or/25-26 (9720)
33    and/31-32 (4)
--------------------------------------------------------------------------------
```

Appendix C.10. Search strategies for HRQoL data

(B): Diverticular disease

```
Database: Ovid MEDLINE(R) ALL <1946 to September 24, 2021>
Search Strategy:
--------------------------------------------------------------------------------
1     *diverticulitis/ (2667)
2     *Diverticulum/ (8202)
3     Diverticul*.mp. (33728)
4     emergency+surgery*.ti,ab. (9470)
5     Drainage*.ti,ab. (97292)
6     Lavage*.ti,ab. (52937)
7     Percutaneous+drainage*.ti,ab. (4232)
8     sigmoidectomy*.ti,ab. (1089)
9     colectomy*.mp. (24998)
10     conservative+manag*.mp. (16769)
11     antibiotic*.ti,ab. (362483)
12     antibiotic+adj+therapy.ti,ab. (0)
13     Anti-Bacterial+Agents/tu (136787)
14     Watchful+wait$.tu. (0)
15     delayed+surg$.ti,ab. (2207)
16     trial.ti,ab. (662690)
17     RCT.ti,ab. (25317)
18     randomi#ed+controlled+trial.pt. (544498)
19     controlled+clinical+trial.pt. (94426)
20     case+control+stud$.ti,ab. (113845)
21     cross-sectional+stud$.ti,ab. (200110)
22     cohort+stud$.ti,ab. (248537)
23     observational+stud$.ti,ab. (128258)
24     Economic+evaluation.ti,ab. (10055)
25     EuroQol-5+Dimension.ti,ab. (690)
26     "EQ-5D".ti,ab. (9680)
27     or/1-3 (10504)
28     or/4-15 (651442)
29     and/27-28 (10504)
30     or/16-24 (1697513)
31     and/29-30 (200)
32     or/25-26 (9957)
33     and/31-32 (2)
--------------------------------------------------------------------------------
```

Appendix C.10. Search strategies for HRQoL data

(C): Abdominal wall hernia

```
-----------------------------------------------------------------------------
1    (inguinal or femoral or ventral or umbilical or abdominal wall).ti,ab. (335807)
2    hernia.ti,ab. (52281)
3    hernioplasty/ (9400)
4    herniorrhaphy/ (9400)
5    hernioplasty.ti,ab. (1602)
6    herniorrhaphy.ti,ab. (2372)
7    repair+or+surg*.ti,ab. (22)
8    hernia+adj+repair.ti,ab. (0)
9    (early adj3 (surg* or repair)).ti,ab. (28362)
10    trial.ti,ab. (657219)
11    RCT.ti,ab. (24987)
12    randomi#ed+controlled+trial.pt. (541163)
13    controlled+clinical+trial.pt. (94345)
14    case+control+stud$.ti,ab. (113048)
15    cross-sectional+stud$.ti,ab. (197037)
16    cohort+stud$.ti,ab. (244943)
17    retrospective+stud$.ti,ab. (179855)
18    observational+stud$.ti,ab. (126477)
19    (cost adj (utility or effectiv*)).ti,ab. (149693)
20    Economic+evaluation.ti,ab. (9983)
21    (quality of life or QoL or HRQoL).ti,ab. (312433)
22    EuroQol.af. (6571)
23    EQ-5D*.af. (9689)
24    and/1-2 (20870)
25    or/3-9 (40553)
26    or/10-20 (1972540)
27    or/21-23 (315435)
28    and/24-27 (129)
-----------------------------------------------------------------------------
```

Appendix C.11. Health-related quality of life (HRQoL) scores from the literature and sources

Condition	**Source**	**Mean age***	**Baseline EQ-5D-3L score**		**One-year EQ-5D-3L score**	
			Females	Males	Females	Males
Acute appendicitis	O'Leary et al., (2021)	32.80	0.751	0.768	0.967	0.989
Diverticular disease	Thornell et al., (2016)	68.00	0.649	0.666	0.866	0.889
Abdominal wall hernia	Rutegård et al., (2018)	58.76	0.848	0.870	0.936	0.960

The three studies used the EuroQol- 5-Dimension (EQ-5D) in its 3-level (3L) version.
*Mean age at trial start in the study.

Appendix C.12. Generalised Linear Models (GLMs) for quality-adjusted life years (QALYs) and costs, assessment of model fit according to root mean squared error (RMSE)

Family	Link	Degree	Acute appendicitis	Diverticular Disease	Abdominal wall hernia
QALYs					
Binomial	Logit	1	[0.059]	[0.192]	[0.204]
Binomial	Logit	2	0.059	0.192	0.204
Binomial	Logit	3	0.059	0.192	0.204
Costs					
Gaussian	Identity	1	3530.330	8980.985	[8975.387]
Inverse gaussian	Identity	1	3533.875	8993.261	8988.748
Gamma	Identity	1	3532.422	8987.824	8983.084
Gaussian	Log	1	3525.947	8977.639	8976.490
Inverse gaussian	Log	1	3530.944	9002.912	9009.775
Poisson	Log	1	3526.854	8980.842	8980.611
Gamma	Log	1	3528.611	8988.305	8990.651
Gaussian	Identity	2	3530.321	8981.013	8975.399
Inverse gaussian	Identity	2	3533.866	8993.318	8988.820
Gamma	Identity	2	3532.425	8987.953	8983.125
Gaussian	Log	2	[3525.900]	[8975.837]	8976.386
Inverse gaussian	Log	2	3530.938	8999.703	9009.797
Gamma	Log	2	3528.617	8985.700	8990.591

Most appropriate GLMs for costs and QALYs were selected looking at RMSEs (in brackets). Degree refers to the polynomial order of the propensity score.

Appendix C.13. Most common operative procedures within emergency surgery window (panel A) and after emergency surgery window and up to one year (B)

(A): Within emergency surgery (ES) window

	Acute appendicitis (N=268,144)		**Diverticular disease (N=138,869)**		**Abdominal wall hernia (N=106,432)**	
	ES (N=247,506)	NES (N=20,638)	ES (N=15,772)	NES (N=123,097)	ES (N=62,559)	NES (N=43,873)
Most common high-volume operative procedures * (%)	Emergency excision of abnormal appendix (63.0) Unspecified other excision of appendix (16.6) Emergency excision of abnormal appendix and drainage (9.9) Unspecified emergency excision of appendix (3.6) Emergency excision of normal appendix (1.3) Other (0.1)	Interval appendicectomy (6.7) Other specified other excision of appendix (4.2) Planned delayed appendicectomy (1.6) Total cholecystectomy (1.6) Image controlled percutaneous drainage of lesion of abdominal cavity (0.4) Other (0.1)	Rectosigmoidectomy and closure of rectal stump and exteriorisation of bowel (55.6) Irrigation of peritoneal cavity (7.7) Sigmoid colectomy and exteriorisation of bowel (6.6) Anterior resection of rectum and exteriorisation of bowel (3.1) Loop colostomy (2.3) Other (14.2)	Fibreoptic endoscopic snare resection of lesion of colon (0.1) Endoscopic division of adhesions of peritoneum (0.1) Endoscopic snare resection of lesion of lower bowel using fibreoptic sigmoidoscope (0.1) Freeing of adhesions of peritoneum (0.1) Fibreoptic endoscopic resection of lesion of colon (0.0) Other (0.0)	Primary repair of inguinal hernia using insert of prosthetic material (27.6) Repair of umbilical hernia using sutures (17.9) Repair of umbilical hernia using insert of prosthetic material (12.3) Primary repair of femoral hernia using sutures (9.4) Primary repair of femoral hernia using insert of prosthetic material (7.0) Other (15.4)	Unspecified urethral catheterisation of bladder (0.5) Ileectomy and anastomosis of ileum to ileum (0.3) Unspecified excision of ileum (0.1) Freeing of adhesions of peritoneum (0.1) Omentectomy (0.0) Other (0.0)

Appendix C.13. Most common operative procedures within emergency surgery window (panel A) and after emergency surgery window and up to one year (B)

(A): Within emergency surgery (ES) window (cont.)

	Acute appendicitis (N=268,144)		**Diverticular disease (N=138,869)**		**Abdominal wall hernia (N=106,432)**	
	ES (N=247,506)	NES (N=20,638)	ES (N=15,772)	NES (N=123,097)	ES (N=62,559)	NES (N=43,873)
% with no 'more common' operative procedures **	4.9	84.9	10.6	99.7	10.5	98.9

Denominator is the total number of patients in the group. *'Other' includes procedures with >1% volume in index admission appearing in ES window. **This Includes patients for whom no procedures were recorded and those who got 'low-volume' (<1%) procedures. NES: non-emergency surgery.

Appendix C.13. Most common operative procedures within emergency surgery window (panel A) and after emergency surgery window and up to one year (B)

(B): After emergency surgery window (ES) up to one year

	Acute appendicitis (N=268,144)		**Diverticular disease (N=138,869)**		**Abdominal wall hernia (N=106,432)**	
	ES (N=247,506)	NES (N=20,638)	ES (N=15,772)	NES (N=123,097)	ES (N=62,559)	NES (N=43,873)
Most common high-volume operative procedures (%) *	Emergency excision of abnormal appendix (2.80) Emergency excision of abnormal appendix (0.78) Emergency excision of abnormal appendix (0.45) Total cholecystectomy (0.31) Unspecified urethral catheterisation of bladder (0.26) Other (1.30)	Unspecified urethral catheterisation of bladder (4.33) Emergency excision of abnormal appendix (4.27) Unspecified other excision of appendix (2.81) Planned delayed appendicectomy (1.88) Planned delayed appendicectomy (0.92) Other (3.98)	Closure of colostomy (9.41) Rectosigmoidectomy and closure of rectal stump and exteriorisation of bowel (7.96) Freeing of adhesions of peritoneum (1.27) Freeing of adhesions of peritoneum (1.14) Sigmoid colectomy and exteriorisation of bowel (1.00) Other (7.23)	Fibreoptic endoscopic snare resection of lesion of colon (4.00) Fibreoptic endoscopic resection of lesion of colon (1.74) Rectosigmoidectomy and closure of rectal stump and exteriorisation of bowel (1.21) Endoscopic snare resection of lesion of lower bowel using fibreoptic sigmoidoscope (1.04) Anterior resection of rectum and anastomosis of colon to rectum using staples (0.60) Other (3.72)	Primary repair of inguinal hernia using insert of prosthetic material (3.83) Repair of umbilical hernia using sutures (1.70) Unspecified urethral catheterisation of bladder (1.52) Repair of umbilical hernia using insert of prosthetic material (1.14) Repair of recurrent inguinal hernia using insert of prosthetic material (0.69) Other (3.01)	Primary repair of inguinal hernia using insert of prosthetic material (26.9) Repair of umbilical hernia using insert of prosthetic material (6.17) Repair of umbilical hernia using sutures (5.56) Repair of recurrent inguinal hernia using insert of prosthetic material (2.86) Unspecified urethral catheterisation of bladder (1.49) Other (6.30)

Appendix C.13. Most common operative procedures within emergency surgery window (panel A) and after emergency surgery window and up to one year (B)

(B): After emergency surgery window (ES) up to one year (cont.)

	Acute appendicitis (N=268,144)		**Diverticular disease (N=138,869)**		**Abdominal wall hernia (N=106,432)**	
	ES (N=247,506)	NES (N=20,638)	ES (N=15,772)	NES (N=123,097)	ES (N=62,559)	NES (N=43,873)
% with no ‘more common’ operative procedures**	94.1	81.8	72.0	87.7	88.1	50.7

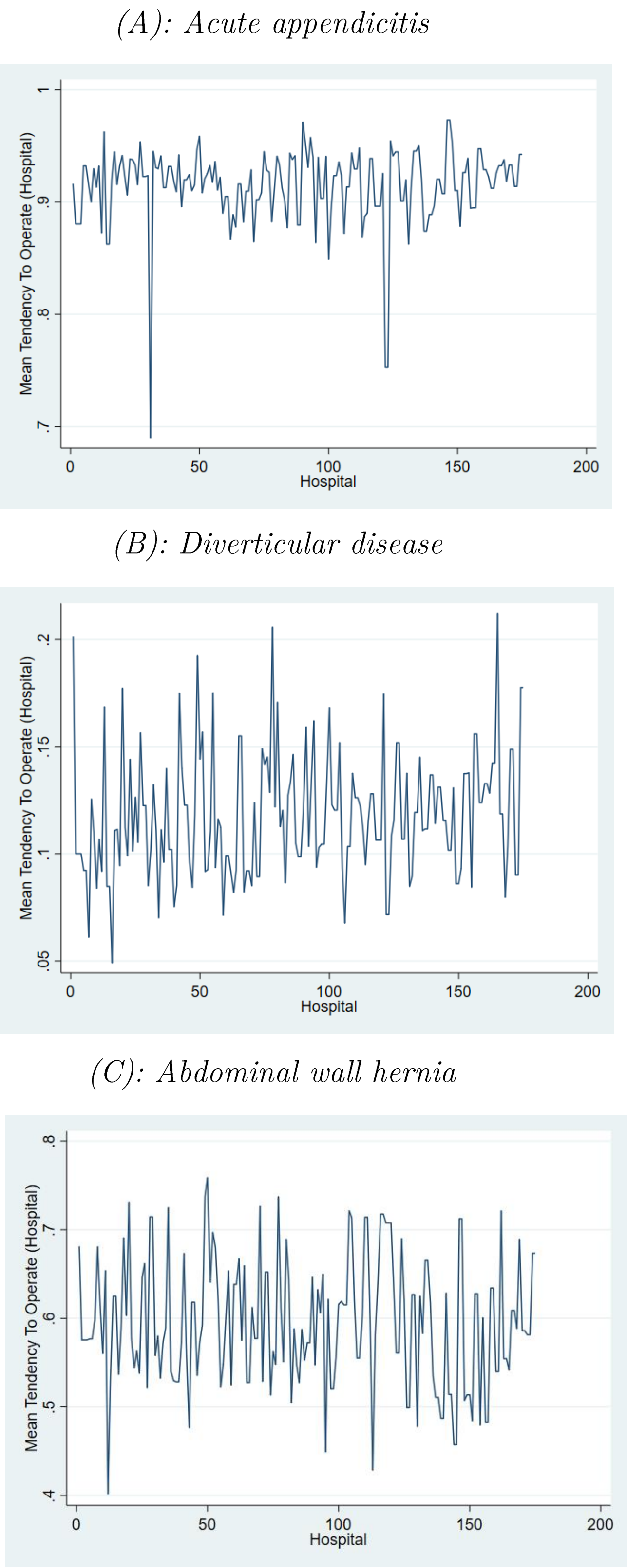
Denominator is the total number of patients in the group. *’Other’ includes procedures with >1% volume in index admission appearing after the ES window. **This Includes patients for whom no procedures were recorded and those who got ‘low-volume’ (<1%) procedures. NES: non-emergency surgery.

Appendix C.14. Summary of sensitivity analyses (SA) results. Overall Incremental Net monetary Benefit (INB) of emergency surgery (ES) vs non-emergency surgery (NES) strategies

Analysis	Description	Acute appendicitis (N=268,144)	Diverticular disease (N=138,869)	Abdominal wall hernia (N=106,432)
Base case	See Chapter 4	-86.2 (-1,163, 991)	2,664 (-4,298, 9,626)	-119 (-1,282, 1,043)
SA1	Considered alternative measures of hospital quality derived from the *National Emergency Laparotomy Audit (NELA)* reports from 2016-2018 (see Section 3.2.4)	408 (-787, 1,605)	5,823 (1,029, 10,616)	125 (-1,027, 1,276)
SA2	Considered a 10% decrease in all unit costs in total cost calculation (see Section 3.2.4)	-96.9 (-1,078, 884)	2,491 (-3,673, 8,655)	-30.4 (-964, 903)
SA3	Considered a 10% increase in all unit costs in total cost calculation (see Section 3.2.4)	-75.2 (-1,257, 1,107)	2,836 (-4,356, 10,028)	-208 (-1,254, 837)
SA4	Used linear interpolation between baseline and one-year HRQOL endpoints for calculating QALYs (see Section 3.2.4)	-202 (-1,514, 1,110)	2796.432 (-2,796, 8,389)	-125 (-1,216, 967)
SA5	Evaluated costs and effects of ES and NES over a five-year time horizon (see Section 3.2.4)	-3,786 (-9,113, 1,541)	-1,502 (-27,066, 24,062)	700 (-6,812, 8,212)

INB of ES in the sensitivity analyses was estimated using Local Instrumental Variable methods. 95% confidence interval in parentheses. ES: emergency surgery, HRQoL: health-related quality of life, NES: non-emergency surgery, QALYs: quality-adjusted life years.

Appendix C.15. Variation in tendency to operate (TTO) across 175 NHS hospitals in the one year prior to emergency admissions that meet the inclusion criteria for acute appendicitis (panel A), diverticular disease (panel B) and abdominal wall hernia (panel C)

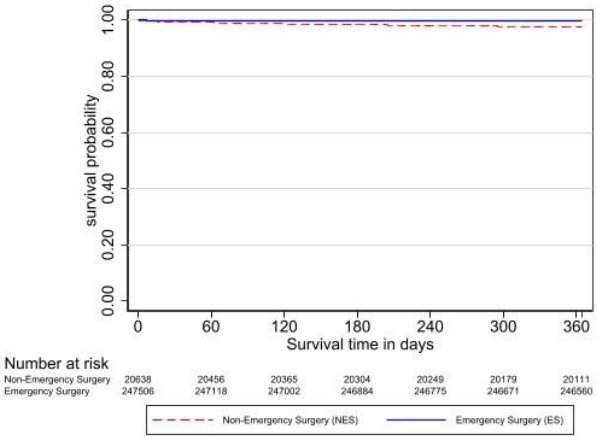

(A): Acute appendicitis

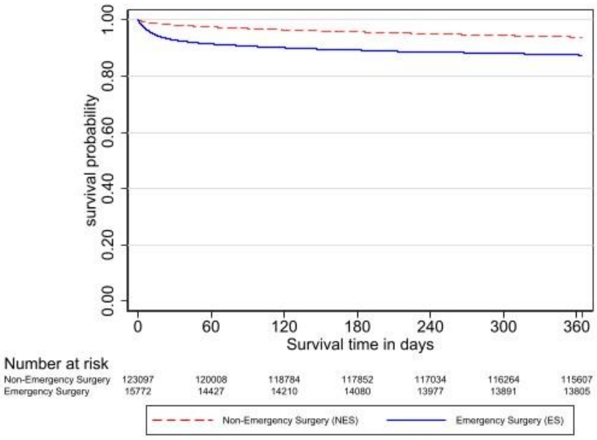
(B): Diverticular disease

(C): Abdominal wall hernia

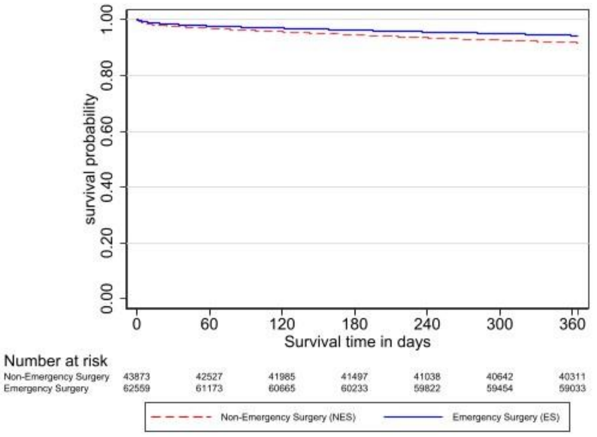
Appendix C.16. Kaplan-Meier estimates for time to one-year death for acute appendicitis (panel A), diverticular disease (B), abdominal wall hernia (C)

(A): Acute appendicitis



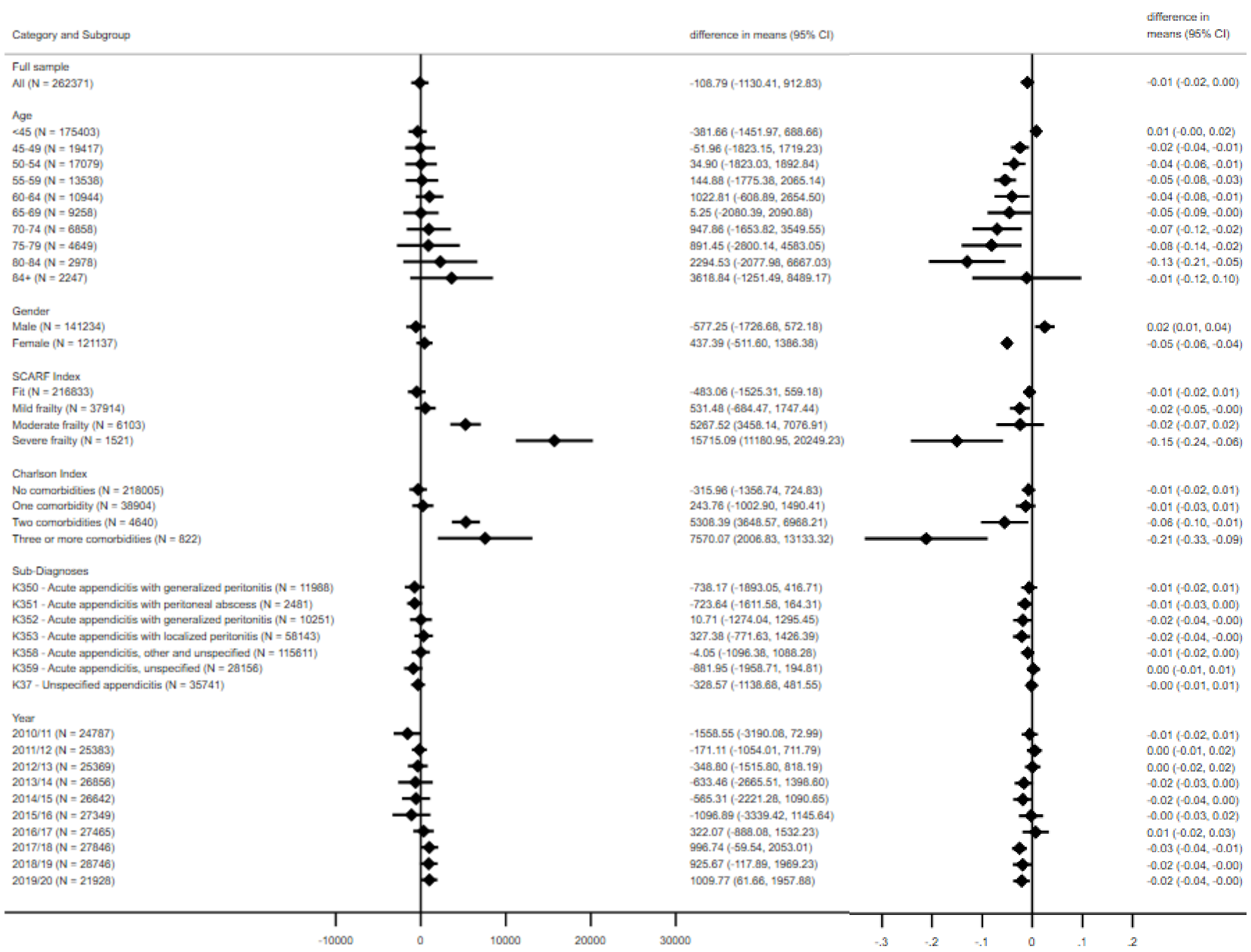


(C): Abdominal wall hernia



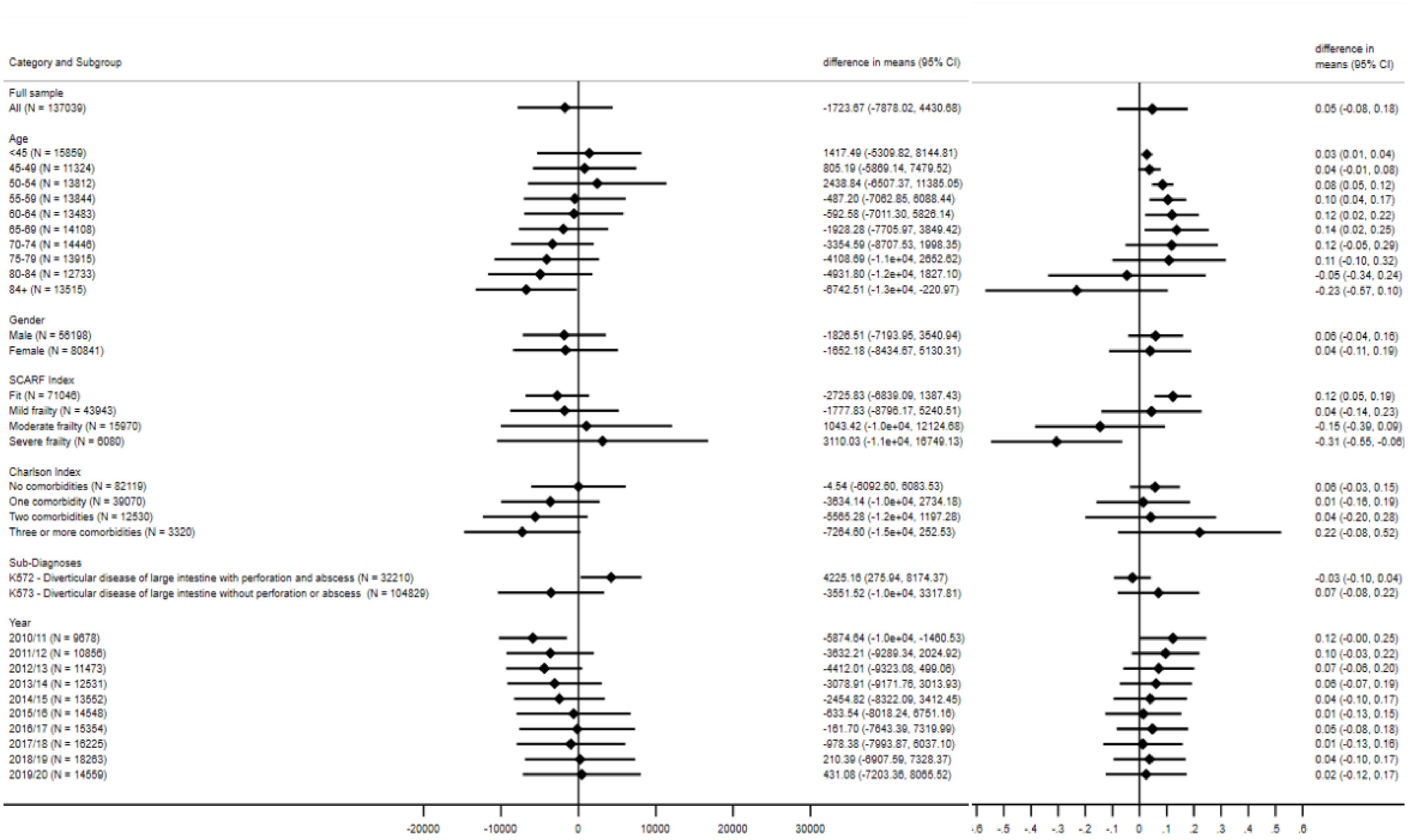
Appendix C.17. Forest plots of estimated incremental costs and Quality-adjusted Life Years (QALYs) from the Local Instrumental Variables (LIV) approach

(A): Acute appendicitis



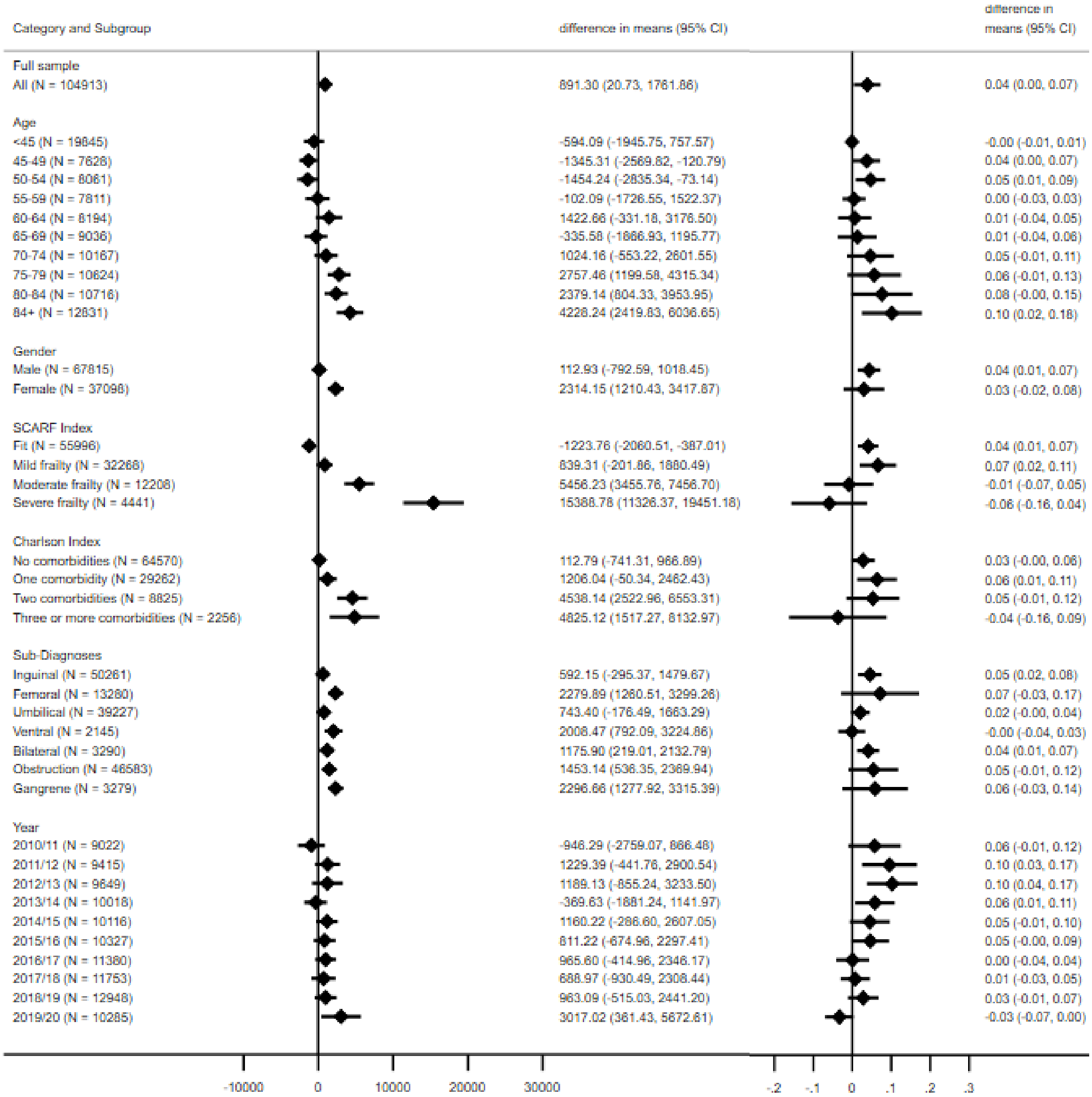
Appendix C.17. Forest plots of estimated incremental costs and Quality-adjusted Life Years (QALYs) from the Local Instrumental Variables (LIV) approach

(B): Diverticular disease

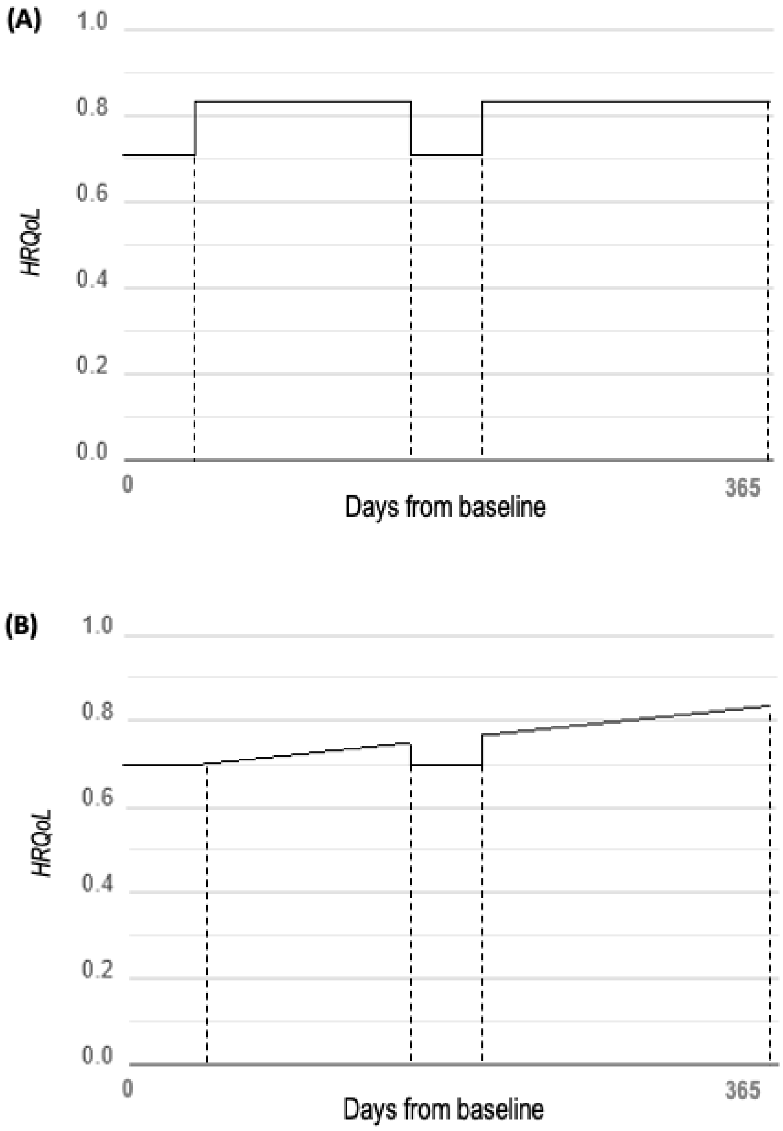


Appendix C.17. Forest plots of estimated incremental costs and Quality-adjusted Life Years (QALYs) from the Local Instrumental Variables (LIV) approach

(C): Abdominal wall hernia



Appendix C.18. Health-related Quality of Life (HRQoL) trajectory following initial (index) emergency admission and emergency readmission for the base case, which assumes HRQoL reaches follow-up levels following hospital discharge (panel A), and linear interpolation (B, sensitivity analysis 4)



(A) Immediate interpolation. Baseline HRQoL is assumed to apply constantly for the duration of the index admission and any emergency readmission. Following the index admission, the HRQoL is assumed to apply constantly for the duration of the period before the final (one-year) endpoint, which is accrued immediately after discharge. (B) Linear interpolation. Baseline HRQoL is assumed to apply constantly for the duration of the index admission and any emergency readmission. HRQoL between the endpoints is assumed to increase linearly.

References

Ara R and Brazier JE (2010) Populating an economic model with health state utility values: Moving toward better practice. *Value in Health* 13(5). International Society for Pharmacoeconomics and Outcomes Research (ISPOR): 509–518. DOI: 10.1111/j.1524-4733.2010.00700.x.

Ara R, Brazier J and Zouraq IA (2017) The Use of Health State Utility Values in Decision Models. *PharmacoEconomics* 35. Springer International Publishing: 77–88. DOI: 10.1007/s40273-017-0550-0.

Basu A (2014) Estimating person-centered treatment (PeT) effects using instrumental variables: an application to evaluating prostate cancer treatments. *JOURNAL OF APPLIED ECONOMETRICS* 29: 671–691. DOI: 10.1002/jae.

Basu A (2015) Person-centered treatment (PeT) effects: Individualized treatment effects using instrumental variables. *The Stata Journal* 15(2): 397–410.

Bjorklund A and Moffitt R (1987) The Estimation of Wage Gains and Welfare Gains in Self-Selection Models. *The Review of Economics and Statistics* 69(1): 42. DOI: 10.2307/1937899.

Clement KD, Emslie K, Maniam P, et al. (2020) What is the Operative Cost of Managing Acute Appendicitis in the NHS: The Impact of Stump Technique and Perioperative Imaging. *World Journal of Surgery* 44(3). Springer International Publishing: 749–754. DOI: 10.1007/s00268-019-05306-2.

Curtis LA and Burns A (2019) *Unit Costs of Health and Social Care 2019 | PSSRU*. Canterbury: Personal Social Services Research Unit, University of Kent.

ESORT Study Group (2020) Emergency Surgery Or NoT (ESORT) study: Study protocol. Available at: https://www.lshtm.ac.uk/media/38711.

Heah S., Eu KW, Ho YH, et al. (1995) Abdominoperineal Resection for Palliation of Advanced Low Rectal Cancer. *Dis Colon Rectum* 5: 1313–17.

Heckman J (1997) Instrumental variables: A study of implicit behavioral assumptions used in making program evaluations. *Journal of Human Resources* 32(3). DOI: 10.2307/146178.

Heckman JJ and Vytlacil E (2005) Structural equations, treatment effects and econometric policy evaluation. *Econometrica* 73(3): 669–738. DOI: 10.1111/j.1468-0262.2005.00594.x.

Heckman JJ and Vytlacil EJ (1999) Local instrumental variables and latent variable models for identifying and bounding treatment effects. *Proceedings of the National Academy of Sciences of the United States of America* 96: 4730–4734. DOI: 10.1073/pnas.96.8.4730.

HM Treasury Department (2020) Gross Domestic Product (GDP) deflators: user guide. Available at: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/205904/GDP_Deflators_User_Guide.pdf (accessed 19 August 2021).

Information Services Division (ISD) Scotland. (2019) Theatres Costs-Detailed Tables - SFR 5.10. Available at: https://www.isdscotland.org/health-topics/hospital-care/operations-and-procedures/.

Ismail I, Wolff S, Gronfier A, et al. (2015) A cost evaluation methodology for surgical technologies. *Surgical Endoscopy* 29(8): 2423–2432. DOI: 10.1007/s00464-014-3929-4.

Javanmard-Emamghissi H, Boyd-Carson H, Hollyman M, et al. (2020) The management of adult appendicitis during the COVID-19 pandemic: an interim analysis of a UK cohort study. *Techniques in Coloproctology* (0123456789). DOI: 10.1007/s10151-020-02297-4.

McCormack K, Wake B, Perez J, et al. (2005) Laparoscopic surgery for inguinal hernia repair: Systematic review of effectiveness and economic evaluation. *Health Technology Assessment* 9(14). DOI: 10.3310/hta9140.

National Emergency Laparotomy Audit (NELA) Project Team (2016) *Second patient report of the National emergency laparotomy audit*. London.

National Emergency Laparotomy Audit (NELA) Project Team (2017) *Third patient report of the National emergency laparotomy audit*. London. Available at: www.nela.org.uk/reports.

National Emergency Laparotomy Audit (NELA) Project Team (2018) *Fourth patient report of the National emergency laparotomy audit*. London.

National Institute for Health and Care (2013) Guide to the methods of technology appraisal. London. Available at: https://www.nice.org.uk/process/pmg9/chapter/foreword.

Neyman J (1990) On the application of probability theory to agricultural experiments. *Statistical Science* 5: 463–480.

NHS Improvement (2018) *NHS reference costs 2017/2018*. London. Available at: https://webarchive.nationalarchives.gov.uk/ukgwa/20200501111106/https://improvement.nhs.uk/resources/reference-costs/.

O'Leary DP, Walsh SM, Bolger J, et al. (2021) A Randomized Clinical Trial Evaluating the Efficacy and Quality of Life of Antibiotic-only Treatment of Acute Uncomplicated Appendicitis: Results of the COMMA Trial. *Annals of surgery* 274(2): 240–247. DOI: 10.1097/SLA.0000000000004785.

Rubin DB (1974) Estimating causal effects of treatment in randomized and nonrandomized studies. *Journal of Educational Psychology* 66(5): 688–701. Available at: http://www.fsb.muohio.edu/lij14/420_paper_Rubin74.pdf.

Rutegård M, Gümüsçü R, Stylianidis G, et al. (2018) Chronic pain, discomfort, quality of life and impact on sex life after open inguinal hernia mesh repair: an expertise-based randomized clinical trial comparing lightweight and heavyweight mesh. *Hernia* 22(3): 411–418. DOI: 10.1007/s10029-018-1734-z.

Sharma P, Boyers D, Scott N, et al. (2015) The clinical effectiveness and cost-effectiveness of open mesh repairs in adults presenting with a clinically diagnosed primary unilateral inguinal hernia who are operated in an elective setting: systematic review and economic evaluation. *Health Technology Assessment* 19(92). DOI: 10.3310/hta19940.

Teramoto A, Aoyama N, Ebisutani C, et al. (2020) Clinical importance of cold polypectomy during the insertion phase in the left side of the colon and rectum: a multicenter randomized controlled trial (PRESECT study). *Gastrointestinal Endoscopy* 91(4). American Society for Gastrointestinal Endoscopy: 917–924. DOI: 10.1016/j.gie.2019.12.019.

Thornell A, Angenete E, Bisgaard T, et al. (2016) Laparoscopic Lavage for Perforated Diverticulitis With Purulent Peritonitis. *Annals of Internal Medicine* 164(3): 137–145. DOI: 10.7326/M15-1210.

Wilson M (2016) Urinary catheterisation in the community: Exploring challenges and solutions. *British Journal of Community Nursing* 21(10). Mark Allen Group: 492–496. DOI: 10.12968/bjcn.2016.21.10.492.

Wu CC, Chueh SC and Tsai YC (2016) Is contralateral exploration justified in endoscopic total extraperitoneal repair of clinical unilateral groin hernias - A Prospective cohort study. *International Journal of Surgery* 36. Elsevier Ltd: 206–211. DOI: 10.1016/j.ijsu.2016.10.012.