

LONDON
SCHOOL of
HYGIENE
& TROPICAL
MEDICINE



LSHTM Research Online

Steventon, A; (2015) Evaluating complex interventions using routinely collected data: Methods to improve the validity of randomised controlled trials and observational studies. PhD thesis, London School of Hygiene & Tropical Medicine. DOI: <https://doi.org/10.17037/PUBS.02212900>

Downloaded from: <https://researchonline.lshtm.ac.uk/id/eprint/2212900/>

DOI: <https://doi.org/10.17037/PUBS.02212900>

Usage Guidelines:

Please refer to usage guidelines at <https://researchonline.lshtm.ac.uk/policies.html> or alternatively contact researchonline@lshtm.ac.uk.

Available under license. To note, 3rd party material is not necessarily covered under this license: <http://creativecommons.org/licenses/by-nc-nd/3.0/>

<https://researchonline.lshtm.ac.uk>

LONDON
SCHOOL *of*
HYGIENE
& TROPICAL
MEDICINE



Evaluating complex interventions using routinely collected data:
Methods to improve the validity of randomised controlled trials
and observational studies

ADAM MARK CHARLES STEVENTON

Thesis submitted in accordance with the requirements for the
degree of:

Doctor of Philosophy

University of London

2015

Department of Health Services Research and Policy

Faculty of Public Health and Policy

LONDON SCHOOL OF HYGIENE & TROPICAL MEDICINE
UNIVERSITY OF LONDON

Funded by: The Nuffield Trust

Research group affiliation: Health Economics, Policy and Technology Assessment Group,
Department of Health Services Research and Policy

I, Adam Mark Charles Steventon, 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.

Abstract

This thesis addresses the evaluation of complex interventions using routinely collected data, specifically the internal validity of observational studies and the generalisability of Randomised Controlled Trials (RCTs). Following a literature review, this thesis has four main objectives: to estimate the effect of telephone health coaching on hospital utilisation in an observational study; to assess optimal choices of control area in observational studies; to estimate the effect of telehealth within a large RCT; and to develop methods to assess aspects of the generalisability of RCTs empirically.

The first paper compares health-coached patients with matched controls. Controls were selected from areas of England that were first matched to the characteristics of the intervention area. Health coaching did not reduce hospital admissions in this study. A second paper uses simulations to assess the relative bias and statistical precision in the treatment effects estimated under alternative approaches to selecting control areas. Lower bias is reported when using local controls than when selecting controls from matched areas, except when there is little unexplained area-level variation in outcomes, when the opposite is true.

The third paper reports that, in the RCT, telehealth patients had fewer hospital admissions than controls, but admissions increased unexpectedly among controls after recruitment, leading to concerns about generalisability. Placebo tests find that control patients in the RCT experienced more admissions than matched non-participants receiving usual care. To address the concern that the control group did not receive 'usual care', sensitivity analyses are presented that contrast outcomes between the telehealth patients in the RCT and matched non-participants. In this comparison, telehealth is associated with a trend towards more admissions than usual care.

The thesis concludes that careful control matching and placebo tests can address important aspects of the validity of observational studies and RCTs, but that further development of evaluation methods is warranted.

Contents

ABSTRACT	3
CONTENTS	4
FIGURES	9
TABLES	10
ACKNOWLEDGEMENTS	12
ABBREVIATIONS	15
CHAPTER 1 INTRODUCTION	17
1.1 COMPLEX INTERVENTIONS.....	17
1.2 THE METHODOLOGICAL CHALLENGES POSED BY COMPLEXITY.....	19
1.2.1 <i>Randomised controlled trials and generalisability</i>	19
1.2.2 <i>The internal validity of observational studies</i>	20
1.3 APPLIED CASE STUDIES.....	21
1.4 ROUTINELY COLLECTED DATA.....	22
1.5 AIMS AND CONTRIBUTIONS OF THIS THESIS.....	24
1.6 CONTRIBUTION OF THE CANDIDATE TO THE THESIS.....	25
1.7 OTHER RELEVANT WORK.....	28
1.8 STRUCTURE OF THE REMAINDER OF THIS DOCUMENT.....	28
CHAPTER 2 CRITICAL REVIEW OF THE METHODOLOGICAL LITERATURE	29
2.1 CONCEPTS AND DEFINITIONS.....	29
2.1.1 <i>Internal and external validity</i>	29
2.1.2 <i>Experiments and observational studies</i>	31
2.1.3 <i>Assignment mechanisms</i>	31
2.2 APPROACHES TO ADDRESS INTERNAL VALIDITY IN OBSERVATIONAL STUDIES.....	32
2.2.1 <i>Regression adjustment</i>	33
2.2.2 <i>Propensity score matching</i>	33
2.2.3 <i>Mahalanobis distance matching on the propensity score</i>	34
2.2.4 <i>Prognostic score matching</i>	35
2.2.5 <i>Genetic matching</i>	36
2.2.6 <i>Combined matching and regression approaches</i>	36
2.3 DESIGN CONSIDERATIONS IN OBSERVATIONAL STUDIES.....	37

2.3.1	<i>Variable selection</i>	37
2.3.2	<i>Source of control patients</i>	38
2.4	THE GENERALISABILITY OF RCT'S	41
2.5	DISCUSSION.....	42
2.5.1	<i>Gaps in the literature regarding the internal validity of observational studies</i>	42
2.5.2	<i>Gaps in the literature about the external validity of RCTs</i>	43
2.6	STRUCTURE OF THE REMAINDER OF THIS THESIS	44
CHAPTER 3 EFFECT OF TELEPHONE HEALTH COACHING (BIRMINGHAM OWNHEALTH) ON HOSPITAL USE AND ASSOCIATED COSTS: COHORT STUDY WITH MATCHED CONTROLS (RESEARCH PAPER 1).....		
45		
3.1	PREAMBLE TO RESEARCH PAPER 1.....	45
3.2	ABSTRACT.....	49
3.3	INTRODUCTION.....	50
3.4	METHODS.....	51
3.4.1	<i>Intervention, including patient recruitment</i>	51
3.4.2	<i>Study populations</i>	53
3.4.3	<i>Study endpoints and sample size calculation</i>	54
3.4.4	<i>Data sources and data linkage</i>	54
3.4.5	<i>Variable definitions</i>	54
3.4.6	<i>Methods to select control group</i>	55
3.4.7	<i>Statistical approach</i>	56
3.4.8	<i>Efforts to avoid bias and sensitivity analysis</i>	56
3.4.9	<i>Ethics approval</i>	57
3.5	RESULTS.....	57
3.5.1	<i>Study populations</i>	57
3.5.2	<i>Comparing hospital use and costs</i>	62
3.5.3	<i>Sensitivity analysis for unobserved confounding</i>	64
3.6	DISCUSSION.....	64
3.6.1	<i>Statement of findings</i>	64
3.6.2	<i>Strengths and weaknesses</i>	64
3.6.3	<i>Comparison with other studies</i>	66
3.6.4	<i>Conclusions</i>	67
CHAPTER 4 COMPARISON OF STRATEGIES TO CHOOSING THE CONTROL POPULATION (RESEARCH PAPER 2).....		
68		
4.1	PREAMBLE TO RESEARCH PAPER 2.....	68

4.2	ABSTRACT.....	71
4.3	INTRODUCTION.....	72
4.4	STATISTICAL CONSIDERATIONS RELATING TO THE CHOICE OF CONTROL AREA.....	74
4.5	CASE STUDY: RAPID RESPONSE SERVICE FOR OLDER PEOPLE.....	77
4.5.1	<i>Case study results</i>	79
4.5.2	<i>Inducing unobserved confounding</i>	79
4.6	DESIGN OF THE SIMULATION STUDY.....	84
4.6.1	<i>Generating baseline data</i>	84
4.6.2	<i>Forming matched control groups</i>	85
4.6.3	<i>Generating outcomes and assessing treatment effects</i>	86
4.6.4	<i>Calibration of the simulation study</i>	86
4.6.5	<i>Scenarios tested</i>	87
4.7	RESULTS OF THE SIMULATION STUDY.....	88
4.7.1	<i>Standardised differences</i>	88
4.7.2	<i>Bias and mean-squared error</i>	91
4.8	DISCUSSION.....	94
4.8.1	<i>Limitations and future research</i>	95
4.8.2	<i>Conclusions</i>	97
4.9	APPENDIX A: CALIBRATION OF THE SIMULATION STUDY.....	99
4.9.1	<i>Individual-level variables</i>	99
4.10	APPENDIX B: SENSITIVITY ANALYSIS.....	101
4.10.1	<i>Matching without replacement</i>	101
4.10.2	<i>Normally distributed outcome</i>	101

CHAPTER 5 EFFECT OF TELEHEALTH ON HOSPITAL USE AND MORTALITY

(RESEARCH PAPER 3)	105	
5.1	PREAMBLE TO RESEARCH PAPER 3.....	105
5.2	ABSTRACT.....	108
5.3	INTRODUCTION.....	109
5.4	METHODS.....	110
5.5	RESULTS.....	116
5.5.1	<i>Data extraction, linkage, and processing</i>	116
5.5.2	<i>Baseline characteristics and trends in hospital activity</i>	117
5.5.3	<i>Analysis of primary and secondary endpoints</i>	120
5.6	DISCUSSION.....	125
5.6.1	<i>Principal findings of the study</i>	125
5.6.2	<i>Strengths and weaknesses of the study</i>	126

5.6.3	<i>Strengths and weaknesses in relation to other studies.....</i>	128
5.6.4	<i>Possible explanations and implications for clinicians and policymakers and other researchers.....</i>	130

CHAPTER 6 AN EMPIRICAL ASSESSMENT OF ASPECTS OF THE GENERALISABILITY

OF THE WHOLE SYSTEMS DEMONSTRATOR TRIAL (RESEARCH PAPER 4)..... 132

6.1	PREAMBLE TO RESEARCH PAPER 4.....	132
6.2	ABSTRACT.....	136
6.3	BACKGROUND	137
6.4	RUNNING EXAMPLE, THE WHOLE SYSTEMS DEMONSTRATOR (WSD) TRIAL.....	139
6.5	CONCERNS ABOUT THE GENERALISABILITY OF THE WSD TRIAL	140
6.6	STATISTICAL METHODS TO ASSESS GENERALISABILITY	142
6.7	APPLYING PLACEBO TESTS TO THE WSD TRIAL	143
6.7.1	<i>Methods.....</i>	143
6.7.2	<i>Results.....</i>	146
6.7.3	<i>Interpretation of the placebo tests.....</i>	150
6.8	SENSITIVITY ANALYSIS REGARDING REASONS FOR NON-GENERALISABILITY	152
6.8.1	<i>Methods.....</i>	152
6.8.2	<i>Results.....</i>	152
6.9	DISCUSSION.....	153
6.10	APPENDIX A: CONFOUNDERS ANALYSED IN PREVIOUS MATCHED CONTROL STUDIES OF TELEHEALTH.....	157
6.11	APPENDIX B: MATCHING AND PLACEBO TESTS.....	157
6.11.1	<i>Numbers of eligible patients.....</i>	157
6.11.2	<i>Generalised linear regression.....</i>	161
6.11.3	<i>Time series analysis.....</i>	162
6.12	APPENDIX C: ANALYSIS FOR TRIAL INTERVENTION GROUP	164

CHAPTER 7 FUTURE DIRECTIONS FOR RESEARCH ON TELEHEALTH (RESEARCH PAPER 5) 169

7.1	PREAMBLE TO RESEARCH PAPER 5.....	169
7.2	INTRODUCTION.....	172
7.3	RESEARCH EVIDENCE	172
7.4	LIMITATIONS OF THE RESEARCH EVIDENCE.....	174
7.5	DEVELOPING BETTER EVALUATIONS.....	176
7.6	MORE SYSTEMATIC DATA COLLECTION.....	177

CHAPTER 8 DISCUSSION 179

8.1	INTRODUCTION.....	179
8.2	FINDINGS FROM THE REVIEW OF THE METHODOLOGICAL LITERATURE.....	179
8.2.1	<i>Improving the internal validity of observational studies.....</i>	<i>179</i>
8.2.2	<i>Assessing and extending the generalisability of RCTs.....</i>	<i>180</i>
8.3	DISCUSSION OF OBJECTIVES RELATING TO OBSERVATIONAL STUDIES.....	180
8.3.1	<i>Applied case study - evaluation of telephone health coaching service.....</i>	<i>180</i>
8.3.2	<i>Implications of the choice of control population.....</i>	<i>182</i>
8.3.3	<i>Contributions of this thesis for the design of observational studies.....</i>	<i>183</i>
8.3.4	<i>Limitations in relation to observational studies.....</i>	<i>183</i>
8.4	DISCUSSION OF OBJECTIVES RELATING TO RCT'S.....	184
8.4.1	<i>Applied case study - evaluation of telehealth.....</i>	<i>184</i>
8.4.2	<i>Placebo tests and associated sensitivity analyses.....</i>	<i>186</i>
8.4.3	<i>Contributions of this thesis for the analysis of RCTs.....</i>	<i>187</i>
8.4.4	<i>Limitations.....</i>	<i>188</i>
8.5	RECOMMENDATIONS FOR POLICYMAKING AND EVALUATION METHODS.....	188
8.6	FUTURE RESEARCH	189
8.6.1	<i>The sufficiency condition for the internal validity of observational studies.....</i>	<i>189</i>
8.6.2	<i>Use and development of placebo tests.....</i>	<i>192</i>
8.6.3	<i>Specifying and testing the logic model.....</i>	<i>193</i>
8.6.4	<i>Formative evaluation.....</i>	<i>194</i>
8.7	CONCLUSIONS	195
	REFERENCES.....	196
	APPENDIX: R CODE FOR SIMULATIONS.....	228

Figures

Figure 1: Trade-offs between internal and external validity.....	32
Figure 2: Numbers of days spent enrolled in telephone health coaching	58
Figure 3: Balance after matching in the telephone health coaching study (selected variables, n=2,698 coached patients and 2,698 matched controls).....	61
Figure 4: Rates of emergency hospital admissions before and after enrolment into telephone health coaching, and equivalent figures for matched controls	63
Figure 5: Area-level variables in the study of the rapid response service, under various strategies for selecting the control population.....	82
Figure 6: Box plots of the estimated treatment effects in the simulation study, under various strategies for selecting the control population.....	93
Figure 7: Box plots of the estimated treatment effects in the simulation study when matching with and without replacement.....	103
Figure 8: Box plots of the estimated treatment effects in the simulation study when using dichotomous and normally distributed endpoints	104
Figure 9: CONSORT diagram showing recruitment into the telehealth study	116
Figure 10: Crude trends in secondary care use for patients recruited into the telehealth study.....	120
Figure 11: Kaplan-Meier survival analysis for the telehealth study.....	124
Figure 12: Patterns of emergency hospital admissions in the telehealth study (n=3,154).....	141
Figure 13: Flow diagram showing numbers of general practices and patients available for the placebo tests.....	144
Figure 14: Crude trends in health care service use for the placebo tests.....	148
Figure 15: Quantile-quantile plots showing balance when applying placebo tests to the RCT control group.....	161
Figure 16: Crude trends in health service use for the sensitivity analysis	167

Tables

Table 1: Balance, before and after matching, in the telephone health coaching study (demographic and health characteristics)	59
Table 2: Balance, before and after matching, in the telephone health coaching study (use of secondary care).....	60
Table 3: Rates of secondary care use before and after enrolment into telephone health coaching, and equivalent figures for matched controls.....	62
Table 4: Estimated treatment effects of telephone health coaching on secondary care use.....	63
Table 5: Standardised differences (%), before and after matching, in the study of the rapid response service, under various strategies for selecting the control population (all individual baseline variables entered into the genetic matching algorithm).....	81
Table 6: Estimated treatment effects of the rapid response service on emergency hospital admissions, under various strategies for selecting the control population.....	83
Table 7: Design of the simulation study, and associations assumed in the base case scenario	84
Table 8: Balance in the simulation study, under various strategies for selecting the control population (observed individual-level confounder).....	89
Table 9: Balance in the simulation study, under various strategies for selecting the control population (unobserved individual-level confounder).....	90
Table 10: Bias (mean-squared error) in the simulation study, under various strategies for selecting the control population.....	92
Table 11: Balance in the simulation study when matching with and without replacement.....	102
Table 12: Bias (mean-squared error) in the simulation study when matching with and without replacement.....	103
Table 13: Baseline characteristics in the telehealth study.....	118
Table 14: Baseline characteristics in the telehealth study (use of secondary care).....	119
Table 15: Rates of secondary care use and mortality during the trial period (unadjusted for clustering and covariates).....	121
Table 16: Estimated treatment effects of telehealth on secondary care use and mortality (results of mixed models, including case-mix adjustment).....	122

Table 17: Balance, before and after matching, when applying placebo tests to the RCT control group (selected variables).....	147
Table 18: Rates of health care use and mortality for RCT control and intervention patients and for their corresponding groups of matched non-participants, during the 12 months following index dates.....	149
Table 19: Results of the placebo tests and sensitivity analysis (generalised linear models).....	150
Table 20: Confounders identified in previous matched control studies of telehealth.....	156
Table 21: Balance, before and after matching, when applying placebo tests to the RCT control group (practice-level variables).....	158
Table 22: Balance, before and after matching, when applying placebo tests to the RCT control group (person-level variables).....	159
Table 23: Balance, before and after matching, when applying placebo tests to the RCT control group (person-level variables, continued).....	160
Table 24: Sensitivity of the placebo test to alternative specifications of the generalised linear models (emergency hospital admissions).....	162
Table 25: Results of the placebo tests (comparison between generalised linear models and time series analysis).....	163
Table 26: Sensitivity of the placebo test to alternative specifications of the time series models (emergency hospital admissions).....	163
Table 27: Balance, before and after matching, in the sensitivity analysis as applied to the RCT intervention group (practice-level variables).....	164
Table 28: Balance, before and after matching, in the sensitivity analysis as applied to the RCT intervention group (person-level variables).....	165
Table 29: Balance, before and after matching, in the sensitivity analysis as applied to the RCT intervention group (person-level variables, continued).....	166
Table 30: Results of the sensitivity analysis (comparison between generalised linear modelling and time series analysis).....	168

Acknowledgements

This thesis could not have been written without patient and constant support from my supervisor, Professor Richard Grieve, to whom I am grateful for our many conversations at coffee shops throughout London and Baltimore. Dr Martin Bardsley was an important collaborator on most of the constituent papers. My partner, Desirée Cooper, always believed that I would finish the project, even though our relationship has never known life without a PhD and there was little empirical evidence to draw on.

Professors Ben Armstrong, Mark Dusheiko and Martin Roland gave valuable comments as members of the advisory group, and Dr Noemi Kreif kindly shared her experience and knowledge. The Nuffield Trust and The Health Foundation supported the work as my employers, and I am particularly grateful to Dr Jennifer Dixon for her advice. The project overlapped with a Harkness Fellowship in Health Care Policy and Practice, and I am grateful for sponsorship from the Commonwealth Fund and to the advice of my Harkness mentor, Professor Harlan Krumholz. Professor Nicholas Barber alerted me to many issues in complexity theory. Below, I give some specific acknowledgements relating to the individual research projects.

For the study of Birmingham OwnHealth (Chapter 3):

Ian Blunt and Dr Martin Bardsley worked with me to design this study. In addition, Sarah Tunkel liaised with the scheme about access to participant data, and Ian Blunt derived unit costs for HES data. I thank the staff in Birmingham East and North Primary Care Trust who organised the data for the health-coached patients; John Grayland for his support on the project; the NHS Information Centre for health and social care for providing invaluable support with the data linkage; and Dr Sally Inglis, Dr Susannah McLean and Professor Douglas Altman for their comments on a previous version of this manuscript.

Funding: This study was funded by the Department of Health in England, which reviewed the study protocol as part of the application for funding and agreed to publication. The views expressed are those of the authors and not of the Department of Health and they do not constitute any form of assurance, legal opinion or advice. The organisations at which the authors were based shall have no liability to any third party in respect to the contents of this paper.

For the comparisons of strategies to select control populations (Chapter 4):

I thank Dr Martin Bardsley, Professor Rhema Vaithianathan and Professor John Billings for some helpful discussions at the early stages of this project. My co-authors, Professors Richard Grieve and

Jasjeet Sekhon, helped to refine the study design and revise the manuscript. Two anonymous peer reviewers gave comments that helped to improve the manuscript greatly.

Funding: This report is independent research supported by the National Institute for Health Research (Senior Research Fellowship, Dr Richard Grieve, SRF-2013-06-016). The views expressed in this publication are those of the author(s) and not necessarily those of the NHS, the National Institute for Health Research or the Department of Health.

For the two studies of the Whole Systems Demonstrator (Chapter 5 and Chapter 6):

I thank the study participants; staff from the three sites; Bupa Health Dialog for managing the remote collection of primary care data and providing predictive modelling analytic data files; and Theo Georgiou and Ian Blunt for their assistance with classifying data for hospital episodes and applying unit costs. The randomised controlled trial was designed and evaluated by the Whole Systems Demonstrator Evaluation Team (principal investigator Professor Stanton Newman). Peer reviewers gave invaluable advice.

For the first paper (Chapter 5), I thank Dr Martin Bardsley, Dr Jennifer Dixon and Professor John Billings, who took part in the original design of this element of the work and contributed to the interpretation of the analysis. Dr Helen Doll was statistical adviser and guarantor of the statistical robustness of the overall project. Dr Martin Cartwright, Dr Shashivadan Hirani, and Dr Lorna Rixon co-ordinated the daily implementation of the trial protocol and maintained trial participants' data. All authors reviewed the manuscript. The Whole System Demonstrator Evaluation Team contributed to periodic discussions of the data collected for this study during team meetings, and commented on interim documents produced during the study.

For the second paper (Chapter 6), I thank my co-authors Professor Richard Grieve and Dr Martin Bardsley, who helped to refine the study design and revise the manuscript.

Funding: The study was funded by the Department of Health in England. The Department of Health reviewed the protocol for the study and provided project manager support for the implementation of telehealth. The University College London Hospitals and University College London were the Whole Systems Demonstrator study sponsors. Their role as sponsors was to ensure that the study was conducted in accordance with the Research Governance Framework for Health and Social Care (Second Edition, April 2005) and to confirm that arrangements were in place for the initiation, management, monitoring, and financing of the trial.

This paper in Chapter 6 is independent research supported by the National Institute for Health Research (Senior Research Fellowship, Dr Richard Grieve, SRF-2013-06-016). The views expressed in this publication are those of the author(s) and not necessarily those of the NHS, the National Institute for Health Research or the Department of Health.

For the paper in Chapter 7:

I thank my co-authors, Dr Martin Bardsley and Professors Brian McKinstry, Richard Grieve, Nicholas Mays, Nicholas Barber and Harlan M. Krumholz for their thoughts and reflections on the paper.

Abbreviations

CI	Confidence Interval
CM	Combined Model
CONSORT	Consolidated Standards of Reporting Trials
COPD	Chronic Obstructive Pulmonary Disease
GP	General Practice
HES	Hospital Episode Statistics
HIV	Human Immunodeficiency Virus
MOR	Median Odds Ratio
MRC	Medical Research Council
NHS	National Health Service
PATE	Population Average Treatment Effect
PATT	Population Average Treatment effect for the Treated
POPPs	Partnership for Older People Projects
QOF	Quality Outcomes Framework
RCT	Randomised Controlled Trial
REACT	Randomised Evaluations of Accepted Choices in Treatment
SATC	Sample Average Treatment effect for Control patients
SATE	Sample Average Treatment Effect
SATT	Sample Average Treatment effect for the Treated
SD	Standard Deviation
SUS	Secondary Uses Service
TIA	Transient Ischaemic Attack

WSD

Whole Systems Demonstrator

Chapter 1 Introduction

1.1 Complex interventions

A growing number of people are living with long-term health conditions such as diabetes, heart failure or chronic obstructive pulmonary disease (Wanless et al. 2006). This poses considerable challenges to health systems globally, as people with long-term health conditions can receive care that is fragmentary as well as costly (Wagner 1998). Health care reform would ideally curb the anticipated increases in cost pressures associated with this growing group of people, while improving their health outcomes and experiences of care (Berwick, Nolan, and Whittington 2008). A specific goal of policymakers is often to prevent admissions to hospitals and care homes (NHS England 2014), as these admissions are expensive as well as undesirable from the perspective of the patient (Guo et al. 2015).

Many interventions have been tested to improve care for people with long-term conditions (Purdy 2010). Telehealth involves the remote exchange of medical information (such as blood pressure) between patient and health care professional (Bower et al. 2011), for example to facilitate faster clinician response to deteriorations to health (Chaudhry, Mattera, and Krumholz 2011). Another example is telephone health coaching, which aims to promote healthy behaviours, improve outcomes and reduce hospital admissions by providing support and encouragement to patients during telephone calls (Hutchison and Breckon 2011). Virtual Wards seek to achieve similar improvements in outcomes through the integration of health and social care for patients who are at risk of adverse events. Virtual Wards use a shared medical record, multidisciplinary team meetings, and automated alert systems that inform community-based health care workers when patients in their care attend emergency departments (Lewis 2006).

These interventions share a common attribute, namely their complexity, but this can be difficult to define. The Medical Research Council's guidance on evaluating complex interventions broadly defines these as having several interacting elements (Craig et al. 2008). Other authors have attempted to distinguish between 'simple', 'complicated' or 'complex' systems (Glouberman and Zimmernan 2002; Hargreaves 2014). In this framework, simple systems are characterised by fixed patterns of behaviour, combined with linear cause-and-effect relationships between system parts. An example from everyday life is following a recipe while preparing an evening meal. Complicated systems are characterised as containing several simple systems, which are not reducible to these systems. For example, leaders might plan and coordinate the activities of multiple teams when sending a rocket to the moon. In contrast, complex systems are characterised by massively entangled webs of relationships that are adaptive as actors learn, coevolve and respond to changes in their environments (*e.g.*, raising a child). Although often used, Glouberman and Zimmernan's framework has been

criticised because it can lead to an under-recognition of the importance of complexity in evaluation, particularly as even apparently simple interventions in health care are highly social processes (Mowles 2014).¹

Regardless of how complexity is defined (Snowden and Boone 2007), the following attributes of telehealth mean that it is likely to be regarded as a complex intervention. First, the objectives of a telehealth programme are often set at a macro level within a local health care economy (*e.g.*, by a commissioner of health services²) but implementation is done at the micro level (*e.g.*, by individual clinics) (Nelson, Batalden, and Godfrey 2007). This means that the details of a telehealth programme are influenced by perceptions about the usefulness of telehealth at both the micro and macro levels, by priorities and resource constraints at both levels, and by the formal and informal relationships between levels. Second, factors external to the local health care economy may influence priorities and resources in ways that affect the development of telehealth services. Examples are Government initiatives to spread telehealth in England (Department of Health 2011) and new financial incentives on health care providers for its use (NHS Commissioning Board 2013). Third, successful implementation of telehealth requires coordinated changes in behaviour from several groups of people, including patients. Regardless of how well specified technological-based interventions are, their effects depend fundamentally on how health care professionals, patients, and family members interact with the technology (*i.e.*, on the 'sociotechnical system'), which is largely unpredictable (Cherns 1976; Wherton et al. 2012). Finally, telehealth programmes aim for multiple outcomes, for example, reduced hospital admissions, improved disease control, improved patient activation, and improved quality of life (McLean et al. 2013).

The complexity of telehealth means that, even if there is broad agreement about the goals of the particular intervention, people might not agree on how telehealth should achieve these goals - *i.e.*, on the logic model (Parry et al. 2013). Thus, some authors have argued that the value of telehealth is to enable faster clinician responses to deteriorations in health, for example, to enable prompt titration of diuretic drugs at the first sign of decompensation in heart failure (Chaudhry, Mattera, and Krumholz 2011). However, other telehealth services have focussed on fostering self-management (Klug et al. 2011) or providing peer-support (Fursse et al. 2008).

¹This was not Glouberman and Zimmerman's intention, as their monograph aimed to elucidate why it is important to recognise complexity in discussing health care reform.

²An organisation with responsibility for arranging or procuring care for a population of individuals.

1.2 The methodological challenges posed by complexity

A central aim of evaluations of complex interventions is to estimate treatment effects,³ whether in the context of a Randomised Controlled Trial (RCT) or in an observational study. The nature of complex interventions can pose many challenges to evaluators in either setting (Craig et al. 2008), particularly in relation to the generalisability of RCTs and the internal validity of observational studies.

1.2.1 *Randomised controlled trials and generalisability*

In RCTs, treatment allocations are made at random, for example assigning some patients to the treatment and others to the control. RCTs have long been favoured in clinical research due to their ability to produce treatment groups that are balanced on both observed and unobserved characteristics at baseline (Bradford Hill 1937). This property gives well-designed and well-conducted RCTs their high 'internal validity'. In other words, provided that the randomisation is properly executed and the sample size is sufficiently large, differences between the outcomes of people assigned to the treatment versus control can be reliably attributed to the interventions under investigation rather than to baseline differences between these groups (Deeks et al. 2003).

Despite the benefits that randomisation brings, the findings from an RCT might not be generalisable beyond the given study context (Roithwell 2005). Problems with generalisability can occur even in RCTs of relatively simple interventions, such as pesticides (Splawa-Neyman, Dabrowska, and Speed 1990). One problem is that the farms selected for such a trial might not be representative of those that would use the pesticide routinely. If farms in the trial and routine practice settings differ in terms of characteristics that modify the treatment effect, then the treatment effect estimated from the RCT data will not apply directly to routine practice. For example, if the trial mostly consists of large farms, then its findings might not apply to groups of small farms. Another problem is that, irrespective of the characteristics of the farms, the treatments may differ between RCT and routine practice settings. For example, the pesticide might be applied in a more regimented way within the context of the RCT than during routine practice. These concerns have been raised in discussions about the distinctions between *efficacy* versus *effectiveness*, in other words to the difference between treatment effects under ideal, laboratory conditions and those under routine practice (Cochrane 1972; Haynes 1999). The interests of policy makers tend to be aligned more closely with questions about effectiveness than efficacy (Bower et al. 2011).

If RCTs can be challenging even for simple interventions such as pesticides, then concerns about their generalisability become only more acute as the interventions become more complex (Flather, Delahunty, and Collinson 2006). This is partly because the interventions themselves may not be as easily replicable (Dixon-Woods et al. 2013). RCT protocols might also constrain the design and implementation of the intervention, limiting the ability of local teams to meet their overarching aims

³ 'Treatment' is used interchangeably with 'intervention' in this thesis.

or to react to changes in the external environment (Hendy et al. 2012). The mere existence of an RCT might affect the behaviour of the people implementing the interventions, since participation may come with actual or perceived political pressure, and represent opportunities or threats to personal and organisational reputations (Greenhalgh 2012; Ettelt, Mays, and Allen 2014). Patients might also act differently when they are aware that they are being studied (Franke and Kaul 1978; McCarney et al. 2007; Marcus et al. 2012). A final challenge arises because patients with long-term health conditions often cannot be blinded to treatment allocations in RCTs of complex interventions. This is because the control treatments (*e.g.*, usual care) are typically already known to these patients and the new interventions typically require changes in their behaviour (Roland and Torgerson 1998). The same argument applies to health care professionals. As a result, both groups might react to knowledge of the treatment allocations in ways atypical of routine clinical practice (Rosnow and Rosenthal 1997). In particular, the literature on RCTs has hypothesised that some patients might be motivated to join RCTs in the hopes of receiving promising new interventions, and moreover the decision to join an RCT might coincide with decisions to change self-care behaviours. As a result, some patients might react negatively when they realise they have been allocated to usual care, a phenomenon termed 'resentful demoralisation' (Cook and Campbell 1979).⁴ Changes in self-care behaviours have been detected in these instances (McCambridge et al. 2014).

These concerns illustrate that, when evaluating complex interventions, the treatments tested in the RCT might be different to those in routine practice, as might the characteristics of the individuals recruited and the wider context (Bate et al. 2014). As a result, trials might not provide accurate estimates of relative treatment effects that apply to routine practice. Efforts to address problems about the generalisability of RCTs appear to have been focussed on the design stage of these trials (Schwartz and Lellouch 1967; Zelen 1990; Olschewski and Scheurlen 1985; Brewin and Bradley 1989), and methods to assess the achieved level of generalisability empirically have not been sufficiently well developed (Cole and Stuart 2010). An important contribution of this thesis will be to illustrate the design and analysis of an RCT of telehealth and to exemplify an approach to assessing its generalisability using routinely collected data.

1.2.2 The internal validity of observational studies

Complexity is also a challenge when conducting observational studies. Unlike RCTs, in which researchers determine the method of treatment allocation, in observational studies the allocations are made externally to the research team, for example based on the preferences of health care professionals and patients during the course of routine clinical practice. Observational studies have the potential to include a broad range of patients, settings and treatment options, representative of those seen in routine clinical practice. However, the internal validity of observational studies can be

⁴ Related issues have been discussed in the literature about job training programmes. These can function as a form of job search for many of their participants, and labour market dynamics can drive the participation process (Heckman and Smith 1999).

threatened when the groups of patients receiving the intervention and control treatments differ at baseline in terms of a characteristic that is also associated with outcome (a 'confounder') (Cochran and Rubin 1973; von Elm et al. 2007).

Theoretically, dealing with confounding is straightforward if the researcher knows the treatment assignment mechanism that determined which patients received which treatment, and moreover has access to the data required to model this mechanism adequately (Rosenbaum and Rubin 1983), or in other restrictive situations (Newhouse and McClellan 1998).⁵ Unfortunately, in almost any practical situation, the nonrandomised nature of the study means that the treatment assignment mechanism is *a priori* unknown. Understanding the treatment assignment mechanism is particularly challenging when the interventions of interest are complex, as in these cases the treatment assignment mechanism is a product of a dynamic and unpredictable system. It might also reflect decisions made at several levels. For example, the eligibility criteria for the treatment of interest might be set at the macro level, but treatment decisions depend additionally on the preferences of individual patients and clinical teams (Freund et al. 2011).

Unlike RCTs, efforts to improve the validity of observational studies have tended to focus on analytical methods, such as how regression models should be specified, rather than on issues of design, such as how the data sets should be assembled prior to an analytical method being applied (Rubin 2008, 2007, 2010). This thesis will illustrate the design and analysis of an observational study of telephone health coaching using routinely collected data. It will also address a design issue that has received scant attention, namely the effect of the choice of control population on the bias and precision of the resulting treatment effects.

1.3 Applied case studies

The methodological contributions of this thesis are motivated and informed by two large case studies: an observational study of telephone health coaching and an RCT of telehealth. In both studies, the primary outcome of interest relates to hospital admissions, since reducing these is an important goal of policymakers.

Previous evaluative studies of telephone health coaching have been of doubtful generalisability (Hutchison and Breckon 2011). Although one large RCT (Wennberg et al. 2010) reported that the intervention was associated with reductions in hospital use, there has been some debate about whether this effect was attributable to the health coaching element of the intervention or to another

⁵ Some of the methods discussed in this thesis, such as prognostic score matching (Section 2.2.4), require knowledge of the relationship between the relevant baseline variables and the outcome, rather than the relationship between the relevant baseline variables and treatment assignment, but the same broad argument applies.

element that related to shared decision making for preference-sensitive conditions (Motheral 2011).⁶ Because there has been continued uncertainty about how to design a successful telephone health coaching service (Hill, Richardson, and Skouteris 2014), this thesis includes a large observational study of a well-established service, namely Birmingham OwnHealth. A significant challenge to the conduct of the study was the need to obtain retrospective data on the health care needs and utilisation of over 3,000 health-coached patients. Also, a control group was needed to allow for the tendency of patients with a history of hospital admissions to show reductions in these admissions over time ('regression to the mean') (Roland et al. 2005).

Evaluations of telehealth have historically been of poor quality (Barlow et al. 2007b; Bergmo 2009). The Department of Health proposed the Whole Systems Demonstrator (WSD) study in 2006 in order to address the concerns about the robustness of the existing evidence base for telehealth (Department of Health 2006). At the time, WSD was the largest trial of telehealth conducted, with over 3,000 participants recruited from community settings in three areas in England. It embodies many of the challenges to generalisability faced when evaluating complex interventions within RCTs, including those associated with non-blinded trials and context-dependent treatments.

1.4 Routinely collected data

Both case studies exemplify approaches to the analysis of routinely collected data - the large amounts of electronic information that are generated as part of the routine operation of health care systems (Black, Barker, and Payne 2004).⁷ The potential of these data to describe the evolution of health need and outcomes for populations has long been recognised (Acheson and Evans 1964), but the more recent digitalisation of health care records has increased the quantities of data available for research (Pope et al. 2014; Stuart et al. 2013; Steventon 2013). The single-payer nature of the National Health Service in the United Kingdom means that data sets can be assembled at the population level across providers and across time. Exploiting these data in evaluations of comparative effectiveness has been identified as a national priority in England (NHS England 2014), with new research centres and infrastructures being established for this purpose (Woollard 2014).

A general concern about the use of routine data in research is that the nature of the data is not under the investigators' control (Roos et al. 1993; Wallihan, Stump, and Callahan 1999). Moreover, the investigators have no control over what information is sought from patients (Hripcsak and Albers 2013; Wennberg et al. 2014). There are also important considerations about individuals' rights,

⁶ Preference-sensitive conditions are those conditions for which at least two valid alternative treatment strategies are available. Since the risks and benefits of the options differ, the choice of treatment involves trade-offs, which in theory should depend on informed patients' preferences (Wennberg, Fisher, and Skinner 2002). An example is with regards to treatment options for arthritis of the hip.

⁷ Routinely collected data include: administrative data sets, such as might be used to generate reimbursement to providers of care (Roos et al. 1993); clinical data that are generated at the point of care, such as the electronic medical record (Sheikh et al. 2011); clinical audit databases (Yudkin and Redman 1990); and patient-generated data (Nelson et al. 2015).

consent and ownership, and for these reasons access to routine data for the purposes of research is protected by certain controls (Bradley et al. 2010). Recent work has used pseudonymisation to link data sets from across primary, secondary and social care without the need for researchers to access patient identifiable information (Bardsley et al. 2011).⁸ Pseudonymisation is used throughout this thesis. The telephone health coaching study used a method developed in previous work (Steventon et al. 2012b) to link a cohort of patients receiving this out-of-hospital intervention to the Hospital Episode Statistics (HES), a national administrative database containing details of all NHS-funded hospital visits in England. The WSD trial also relied on routine data sets for its primary analyses, because of the difficulty and expense of obtaining self-reported data from all 3,000 participants (Newman et al. 2014). My work on this trial required extracting and linking person-level data from four commissioners of hospital care, over 250 general practices, three local authorities, and several community health services. Over one billion rows of data were extracted in total, making it a novel and extensive example of data linkage within an RCT setting.

This thesis contributes to the methodological toolkit for the analysis of routine data, by developing methods that exploit their breadth (*i.e.*, number of patients) and length (*i.e.*, number of years covered) in order to increase aspects of the validity of RCTs and observational studies. A particular challenge when using routinely collected data in observational studies relates to the ability to model the treatment allocation mechanism (Rosenbaum and Rubin 1983), since, even if this mechanism is well understood, data might still be missing on some confounders. Approaches to addressing this unobserved confounding include: using clinically-generated data, rather than administrative data (Austin et al. 2005); applying difference-in-difference methods (Stuart et al. 2014); and conducting sensitivity analyses for unobserved confounding (Rosenbaum 2002). However, few studies have sought to exploit the fact that routine data often span multiple health care centres and geographic areas (Padkin, Rowan, and Black 2001). Thus, several nonrandomised control groups are often available with which to make comparisons. For example, if an intervention is being tested within a certain general practice, then controls might be taken from within the same general practice as intervention patients, from other, similar general practices, or nationally. As we will see, the choice has implications for unobserved confounding. These have rarely been assessed (Stuart and Rubin 2008), even though studies have used the full range of approaches to selecting control groups (McConnell et al. 2008; Nelson 2012; Roland et al. 2012).

Routinely collected data contain rich information on patients who do not participate in the RCTs, yet may receive the treatments of interest in the course of routine practice. Therefore, various authors have proposed using routine data sets to assess the generalisability of RCTs by means of comparisons

⁸ Pseudonymisation is a method of de-identifying data sets, which involves the removal of patient identifiable fields (such as name or address) and the encryption of a unique patient identifier (such as the NHS number in England). Data sets are linked using the encrypted version of the unique identifier. Pseudonymisation is currently accepted by the English National Health Service as a valid approach to data linkage provided that sufficient other controls are in place (Department of Health 2014).

between the outcomes of people who were recruited into an RCT and those who received the same treatment in routine practice (Lewsey et al. 2000; Padkin, Rowan, and Black 2001).⁹ The potential of routinely collected data to assess generalisability is particularly apparent for certain trial designs. For example, point-of-care trials aim to recruit patients within a regular clinical consultation by using the electronic medical record to alert doctors to the presence of an eligible patient, to record randomly-generated treatment allocations, and to obtain information about outcomes (van Staa et al. 2014; Fiore et al. 2011). Despite the potential for using routinely collected data to assess generalisability, few analytical methods have been developed for this purpose (Cole and Stuart 2010).

1.5 Aims and contributions of this thesis

The overall aim of this thesis is to improve the estimation of treatment effects when evaluating complex interventions for people with long-term health conditions using routinely collected data. The thesis makes both applied and methodological contributions. First, I show how the design, analysis and interpretation of an important observational study of telephone health coaching can be done using routinely collected data, using a matching method and sensitivity analyses for unobserved confounding. Second, I focus on a specific aspect of the design of observational studies that has not been sufficiently addressed in the existing literature (Stuart and Rubin 2008), namely the effect of strategies for choosing the control population on the bias and precision of the resulting estimates of treatment effect. This emerged as an important consideration from the telephone health coaching study, and it will be relevant to many analyses of routinely collected data.

The third and fourth contributions relate to RCTs. I begin by demonstrating how routine data can be obtained and linked at the population level within the context of a multi-centre RCT. Then, I demonstrate an approach that uses linked, routinely collected data to estimate the effect of telehealth within this RCT. I extend analytical approaches to assessing the generalisability of RCTs (Cole and Stuart 2010; Hartman et al. 2015) with new sensitivity analyses that can be conducted when the comparisons fail to confirm the generalisability of the original study. Finally, a policy paper draws together the implications for policymakers and proposes a way to use routinely collected data in formative evaluations of complex interventions (Scriven 1966), with a specific focus on the logic model. The use of formative evaluation designs was not an objective of this thesis, but the findings of the case studies motivate their use in future work.

⁹ Some authors have preferred prospective, bespoke approaches to data collection for nonrandomised patients (Olschewski, Schumacher, and Davis 1992), but this can be costly to implement in practice.

The specific methodological and applied objectives of this thesis are:

Methods review

1. To critically review methods appropriate to addressing the internal validity of observational studies of complex interventions and the external validity of RCTs of these interventions (Chapter 2).

Observational studies

2. To estimate the effect of Birmingham OwnHealth on the utilisation of hospitals, by employing an observational study design (Chapter 3).
3. To assess the implications of the choice of control population in observational studies for the bias and efficiency of the resulting estimates of relative treatment effect (Chapter 4).

Randomised controlled trials

4. To estimate the effect of telehealth on hospital utilisation and mortality, within a large RCT (Chapter 5).
5. To develop an analytical method to assess the generalisability of RCTs, together with a sensitivity analysis to explore the consequences of potential causes of non-generalisability (Chapter 6).

Recommendations

6. To make recommendations for policy development and evaluation methods (Chapter 7).

1.6 Contribution of the candidate to the thesis

Besides the methodological review (Objective 1), all objectives have been met through research papers. With the exception of the paper in Chapter 7, these have either been published, or are forthcoming, in peer-reviewed journals. The following paragraphs summarise my contribution to each paper. I would especially like to acknowledge Professor Richard Grieve for his invaluable and supportive supervision throughout. All authors read and approved the draft papers before they were submitted to journals.

Objective 1: Critically review methods appropriate to addressing the internal validity of observational studies of complex interventions and the external validity of RCTs of these interventions

I conducted this literature review under the guidance of Professor Richard Grieve.

Objective 2: Estimate the effect of Birmingham OwnHealth in an observational study

This paper has been published in the *BMJ* (Steventon et al. 2013b). Chapter 3 contains the published version of the text.¹⁰ This work was conducted between April 2011 and June 2013.

My contribution included the overall design of the study, including the approach to matching, estimation and sensitivity analysis. I liaised with Birmingham East and North Primary Care Trust to arrange for the transfer of identifiable data to the NHS Information Centre for health and social care for data linkage. I conducted the analyses and produced the first draft of the manuscript, before incorporating comments from other co-authors and from peer reviewers.

Objective 3: Assess the implications of the choice of control population in observational studies for the bias and efficiency of the resulting estimates of relative treatment effect

This paper has been published in *Health Services Research & Outcomes Methodology* (Steventon, Grieve, and Sekhon 2015). Chapter 4 contains the version that was accepted for publication. This work was conducted between January 2011 and December 2014.

The paper originated from the candidate, following discussions with Dr Martin Bardsley and Professor John Billings about the optimal choice of control population in observational studies. I contributed to the specification of the research question and proposed the study design for the case study analysis, which Professor Richard Grieve helped refine. All authors contributed to the design of the simulation study. I conducted all analyses, including the simulations, and produced the first draft of the manuscript, before incorporating comments from the co-authors and peer reviewers.

Objective 4: Estimate the effect of telehealth on hospital utilisation in a large RCT

This paper has been published in the *BMJ* (Steventon et al. 2012a). Chapter 5 contains the published version. This work was conducted between November 2008 and June 2012. Data collection went from November 2008 to December 2010, and the analyses were conducted from December 2010 onwards.

A team led by principal investigator Professor Stanton Newman designed the Whole Systems Demonstrator trial, which is registered with the International Standard Randomised Controlled Trial Number Register (ISRCTN43002091). I contributed to the final protocol for the trial (Bower et al. 2011) and led the data collection and analysis for 'Theme One' of the evaluation, which concerned impacts of telehealth on health care utilisation and mortality. Dr Martin Bardsley, Dr Jennifer Dixon, Dr Helen Doll and Professor John Billings were the other members of the Theme One evaluation team and we met to discuss the project on a regular basis. My specific roles included arranging for

¹⁰ All of the papers have been reformatted to meet the stylistic requirements of this document. Some minor amendments have been made to the texts (for example, to convert American spelling into English spelling).

trial participants to be linked to national Hospital Episode Statistics (HES) data, and supervising a subcontractor (Bupa Health Dialog), who provided the trial sites with technical assistance for the extraction of data from general practices and the Secondary Uses Service (SUS).¹¹ Bupa Health Dialog structured the SUS and GP data, and calculated predictive risk scores, which I then verified. I collaborated with the trial statistician (Dr Helen Doll) and representatives from the other themes in designing the details of the analytical approach, including the specification of the regression models and the presentation of the data. I conducted all analyses and produced the first draft of the manuscript, before incorporating comments from other co-authors and peer reviewers.

Objective 5: Develop an analytical method to assess the generalisability of RCTs, together with a sensitivity analysis to explore the consequences of potential causes of non-generalisability

This paper was accepted by *Medical Decision Making* in February 2015 and is forthcoming. Chapter 6 contains a copy of the manuscript after the incorporation of the most recent set of peer-reviewer comments. This work was conducted between January 2011 and January 2015.

The original funding proposal for the WSD evaluation (prepared by Professor Stanton Newman, Dr Jennifer Dixon, Professor John Billings and others) specified risk-adjusted comparisons between the hospitalisation outcomes of RCT intervention patients and those of patients living in other areas of England. I worked with Professor Richard Grieve to design a study that used placebo tests to examine aspects of generalisability more directly, and together we developed the sensitivity analysis that is included in the paper. I structured the data sets, conducted the matching and statistical analyses and produced the first draft of the manuscript, before incorporating comments from the co-authors and peer reviewers.

Objective 6: Make recommendations for policy development and evaluation methods

This paper has not yet been submitted for publication. It was prepared during 2013 and the complete author list is: Adam Steventon, Martin Bardsley, Brian McKinstry, Richard Grieve, Nicholas Mays, Nicholas Barber, and Harlan M. Krumholz. This paper was the suggestion of co-author Nicholas Mays. I conducted most of the background research and prepared the first draft of the manuscript. Co-authors then contributed thoughts and reflections, which I reflected in the manuscript.

¹¹ SUS is an alternative source of hospital data. It was preferred to HES for some of the analyses for technical reasons relating to the data linkage methods. The two data sets were found to be very similar.

1.7 Other relevant work

During the time frame of this thesis, I have also contributed to a number of other observational studies (Steventon et al. 2012b; Roland et al. 2012; Chitnis et al. 2013) and analyses of RCTs (Steventon et al. 2013a, 2014; Bardsley, Steventon, and Doll 2013; Henderson et al. 2014, 2013; Cartwright et al. 2013; Bower et al. 2011). These papers are not reproduced here but are referenced throughout the thesis.

1.8 Structure of the remainder of this document

Chapter 2 reviews methods appropriate to addressing the internal and external validity of RCTs and observational studies of complex interventions. Chapters 3-7 present the research papers, which are each preceded by a short preamble that puts them into the context of the thesis. Chapter 8 concludes.

Chapter 2 Critical review of the methodological literature

Establishing internal validity and generalisability is critical for any study that aims to inform policymaking by estimating treatment effects. This chapter puts both concerns into the single framework provided by Rubin's potential outcomes theory, before reviewing the main methods that can be used to address them within observational studies and RCTs.

The main threats to validity in observational studies tend to relate to internal validity, while RCTs tend to be more susceptible to problems with generalisability. Because both issues represent different facets of the same problem, namely how to deal with selection bias due to an unknown assignment mechanism, we apply the lessons from RCTs to observational studies, and the lessons from observational studies to RCTs. Thus, I identify gaps in the methodological literature for both types of study, some of which will be addressed by this thesis.

2.1 Concepts and definitions

2.1.1 *Internal and external validity*

Rubin conceptualised each individual as having two or more potential outcomes (Rubin 2008), depending on which treatment they receive at a certain point in time. We focus on the situation in which there are two treatments, known as the 'intervention' (A) and the 'control' (B).¹² The *treatment effect* for a particular individual can be defined as the difference between these potential outcomes:

$$TE_i = Y_i(A) - Y_i(B)$$

The fact that only one of these potential outcomes is observed for each individual at each point in time is known as the 'fundamental problem of causal inference' (Holland 1986). Progress can be made because studies are typically concerned with average treatment effects over groups of individuals, rather than for particular individuals. For example, in estimating treatment effects from an RCT, the observed outcomes of the treatment group are contrasted to those of the control group, and the control group's outcomes are effectively used to 'impute' the potential outcomes of the intervention group under the control treatment.

¹² We refer to patients who received the intervention as 'intervention patients' or 'treated patients'. Likewise, patients who received the control are referred to as 'control patients', 'untreated patients', or simply 'controls'. The phrases 'treatment effect' and 'intervention effect' are also used synonymously.

A common estimand is the *Sample Average Treatment Effect* (SATE) (Imai, King, and Stuart 2008):

$$SATE(A,B) = \frac{1}{n} \sum_i TE_i$$

Here, the summation is over all patients in the sample of individuals included in the study, and n is the sample size. Another estimand is the *Sample Average Treatment effect for the Treated* (SATT), which is defined as:

$$SATT(A,B) = \frac{1}{n_1} \sum_{i,Z_i=1} TE_i$$

Unlike the formula for SATE, this summation only includes people who received the intervention. The *Sample Average Treatment effect for Control patients* (SATC) can be defined similarly as the summation over controls.

Although studies often report sample treatment effects such as SATE, SATT or SATC, often of more relevance to policymakers are population effects, where the population refers to the people who would be affected by the decision that the study findings are intended to inform.¹³ The *Population Average Treatment Effect* (PATE) is defined in a similar way to SATE, but treatment effects are summed over all individuals in the population rather than over individuals in the sample.

Imai and colleagues decomposed the difference between PATE (the assumed target quantity of interest) and D (the estimator) (Imai, King, and Stuart 2008) as follows:

$$PATE - D = \Delta_T + \Delta_S$$

Where $\Delta_T = SATE - D$ and $\Delta_S = PATE - SATE$. The first term relates to *bias from internal validity*, or whether the estimated sample treatment effect reflects the quantity it purports to estimate. The second term relates to *bias from generalisability* (or *external validity*), or whether the treatment effect targeted by the study is the same as the effect for the population of interest. The bias from external validity is zero when the sample is a random survey of the entire population, or when the individual treatment effects are the same across the entire population (*i.e.*, there is no treatment effect heterogeneity). It will in general be nonzero when these criteria do not hold, or when the treatments delivered within the study setting have different effects on outcomes to those in routine practice.

¹³ Although studies may be intended to inform decisions relating to several populations, for simplicity I will assume a single population.

2.1.2 *Experiments and observational studies*

Experiments are defined as studies in which the investigator has control over the treatment allocations (Deeks et al. 2003). In contrast, in an *observational study*, treatment allocations are made independently of the investigator (*e.g.*, based on patient preference). A common example of an experiment is the Randomised Controlled Trial (RCT), in which treatment allocations are made randomly. An example of an observational study occurs when investigators track the outcomes of a cohort of patients who received the treatment of interest within the course of routine clinical practice.

2.1.3 *Assignment mechanisms*

Rubin argued that, in order to overcome the fundamental problem of causal inference, it is necessary to formulate an *assignment mechanism* (Rubin 2008). This describes the reasons for the missing potential outcomes:

$$P(W|X, Y(0), Y(1))$$

Here, W is the treatment assignment, and X is a vector of baseline covariates.

The assignment mechanism is important in both experiments and observational studies. However, in an experiment, it is known by design for patients who are included in the study (Little and Rubin 2000). For example, it is equal to 0.5 in an RCT with two treatment groups and a 1:1 allocation ratio. For observational studies, the assignment mechanism is usually unknown and must be estimated from the data (Rubin 2008). The uncertainty inherent in accounting for this unknown assignment mechanism is what threatens the internal validity of observational studies.

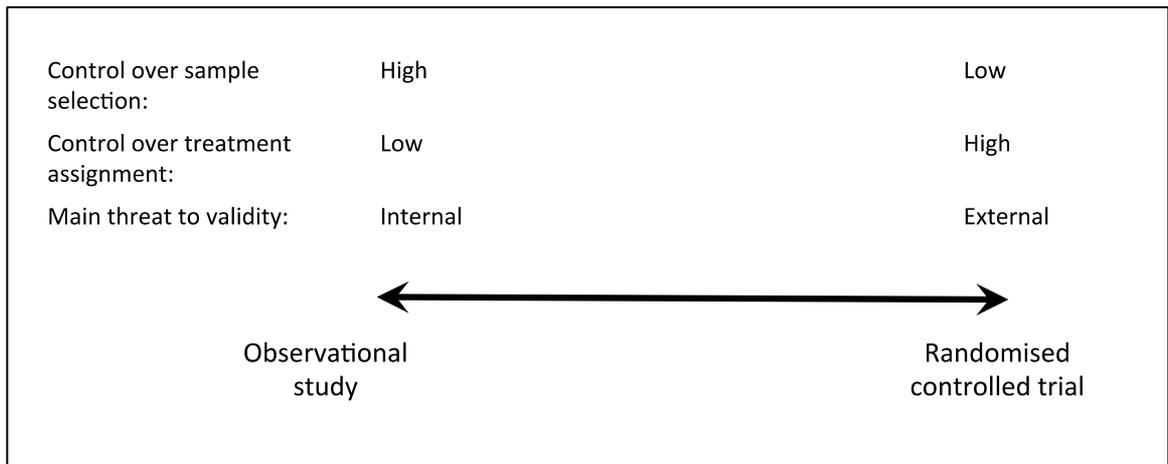
Despite their weaknesses in relation to internal validity, observational studies often have greater control than RCTs over the selection of patients into the study. This is because observational studies can often sample directly from the population that is receiving the treatment in routine practice, while RCTs have to recruit health care organisations, health care professionals, and patients prospectively, and not all will agree to participate.¹⁴ The uncertainty inherent in the sample assignment mechanism is partly what accounts for the threats to generalisability in RCTs.¹⁵ Figure 1 describes the overall situation and the trade-offs that result from sacrificing investigator control over sample selection for greater control over treatment assignments.

The next few sections review approaches to address the internal validity of observational studies. We then turn to approaches to address the generalisability of RCTs.

¹⁴ For a review of some reasons why the characteristics of patients recruited into an RCT may differ from those of patients receiving the treatment in routine practice, see Rothwell (2005). A qualitative study explored some of these issues in the context of the Whole Systems Demonstrator trial, which forms the case study RCT for this thesis (Sanders et al. 2012).

¹⁵ Generalisability can also be threatened when the treatment (or the context in which it is delivered) is different in the RCT than in routine practice.

Figure 1: Trade-offs between internal and external validity.



This figure puts observational studies and RCTs on a spectrum, as some study designs fall between the two (such as a quasi-experimental study in which treatment allocations are based on the order in which participants are recruited). The figure is only illustrative, and it is possible to have, for example, an RCT with low internal and external validity, or an observational study with high internal and external validity.

2.2 Approaches to address internal validity in observational studies

Many methods to estimate treatment effects from observational data require an assumption that is variously called *no unobserved confounding*, *selection on observables*, or *strong ignorability*.¹⁶ Specifically, conditional on the observed baseline covariates, X , the potential outcomes must be independent of treatment assignment:

$$P(W|X, Y(0), Y(1)) = P(W|X)$$

Although some analytical methods exist that do not require this assumption, they can often not be applied in health services research. For example, instrumental variable analysis requires the existence of a variable that is highly correlated with treatment but does not independently affect the outcome (Newhouse and McClellan 1998). Examples of valid instruments in health services research are rare (Stukel et al. 2007). Similarly, regression discontinuity methods can only be applied in unusual instances where there is a variable and a level of that variable such that the probability of receiving the intervention is discontinuous (Imbens and Lemieux 2008).

¹⁶ This is closely related to the assumption called 'no essential heterogeneity', which specifies that selection into treatment should not be a function of the gain from treatment, conditional on observed covariates (Heckman, Urzua, and Vytalil 2006), and so the placebo tests that I go on to describe would also be applicable for assessing this assumption. A confounder is a variable that is associated with both the outcome and treatment assignment (Drake 1993).

As the 'no unobserved confounding' assumption is likely to be needed when conducting observational studies in the current context, I now focus on reviewing methods that use this assumption to estimate treatment effects, namely:

1. Regression adjustment;
2. Propensity score matching;
3. Mahalanobis distance matching on the propensity score;
4. Prognostic matching;
5. Genetic matching; and
6. Combined matching and regression approaches.

The assumption that there is no unobserved confounding cannot be empirically tested. Therefore, in Section 2.3, we will turn to some methods to limit the susceptibility of the analytical strategy to violations of this important assumption.

2.2.1 Regression adjustment

A common analytical approach to estimating treatment effects is to fit parametric models ('regression models') to the data (Harrell, Lee, and Mark 1996). This approach assumes knowledge of the functional form of the relationship between baseline covariates, intervention status and outcome, which must be expressed in the form of a regression equation. Estimates will depend on the functional form assumed, especially when there is poor overlap between the distributions of baseline covariates in the treated and untreated groups (King and Zeng 2006). Unfortunately, the functional form is usually not known *a priori*. Best practice is therefore to conduct sensitivity analysis assuming different combinations of predictors, interaction terms and link functions. One of the benefits of the matching methods that I now describe is that the dependence of the treatment effect estimates on model specification is typically lower if regression models are applied to matched data than to the original sample (Ho et al. 2007).

2.2.2 Propensity score matching

Rosenbaum & Rubin (1983) elucidated the importance of *balancing scores* in causal inference. Balancing scores are variables such that, conditional on these variables, intervention status is independent of the baseline covariates. Rosenbaum and Rubin's central theorem was that, at any level of a balancing score, a comparison of the outcomes of treated and untreated patients will yield an unbiased estimate of the treatment effect at that level of the balancing score. Therefore, if one can obtain, for every treated patient, an untreated patient that is sufficiently closely matched on the balancing score,^{17,18}

¹⁷ For a description of some matching algorithms that can be used, see Austin (2008a).

then the difference between the mean outcomes of the matched groups will yield an unbiased estimate of the sample treatment effect.¹⁹ For this approach to work there must be *overlap* or *positivity* - in other words, there must be no level of the balancing score at which the intervention is received with certainty, or not received with certainty.

The *propensity score* is one example of a balancing score, and it is defined as a individual's probability of receiving the intervention conditional on observed baseline covariates (Rosenbaum and Rubin 1983). In observational studies, the propensity score is usually unknown, and so must be estimated from the data, often using logistic regression.²⁰ Rosenbaum & Rubin (1983) showed that, in finite samples, using consistent estimates of the propensity score produces asymptotically unbiased estimators (*i.e.* bias tends to zero as the sample size increases). However, incorrectly specified propensity score models can lead to bias (Drake 1993). Since there will always be a risk that propensity scores are misspecified, it is important to check *balance* on matched samples and refine the specification of the propensity score model if needed (Austin 2009a) - see Box 1.

2.2.3 Mahalanobis distance matching on the propensity score

Although the theoretical results of Rosenbaum and Rubin (1983) showed that matching on a correctly-specified propensity score will produce balanced groups in expectation, this may not be the case in the context of a particular study, either because the propensity score model is misspecified or because of chance. Therefore, it is often advisable to match not only on the propensity score but also on prognostically important baseline variables by using a multivariable distance measure such as the Mahalanobis distance (Mahalanobis 1936). Rosenbaum and Rubin (1985) found that matching on the Mahalanobis distance produced more closely balanced groups than matching on the propensity score alone.

¹⁸ Matching is not the only way to use a balancing score. For example, it is also possible to stratify the population into broad bands based on the propensity score, to weight by the inverse of the probability of treatment assignment, or to do covariate adjustment using the propensity score. In one set of simulations, matching on the propensity score produced better balance on observed confounders than stratification (Austin, Grootendorst, and Anderson 2007). A subsequent study found that matching on the propensity score and inverse probability of treatment weighting both eliminated a greater degree of the systematic differences between treated and untreated subjects than stratification and covariate adjustment (Austin 2009b). More recent work has found that inverse probability of treatment weighting can lead to unstable estimates due to extreme weights (Radice et al. 2012).

¹⁹ If the matching is conducted in this way, then the treatment effect estimated will be SATT. It is also possible to use matching methods to estimate SATE or SATC (Stuart 2010).

²⁰ Several other methods have been developed for fitting propensity scores, including machine-learning approaches such as classification and regression trees (Lee, Lessler, and Stuart 2010).

Box 1: Assessing balance in matched control studies

Balance has been defined as the similarity between the multivariate empirical distributions of the covariates in the intervention and matched control groups (Imai, King, and Stuart 2008).²¹ There is neither an accepted method for assessing balance nor a common view about what constitutes adequate balance (Hill 2008), but there appears to be consensus that it is inappropriate to compare the distribution of baseline covariates using statistical tests (Ho et al. 2007). Two reasons are given. The first is the observation that balance is a property of the sample and not of a hypothetical population. Second, probabilities given by significance tests are influenced by sample size. This means that a statistical test could conclude that there is a 'non-significant' imbalance simply because of a lack of power to detect differences. There are dissenting views about the value of statistical tests, however (Hansen 2008a). One simulation study showed that a strategy that selects propensity score models based on an assessment of balance using the t-test can result in lower bias than strategies using other metrics such as the standardised difference or the Kolmogorov-Smirnov test statistic (Belitser et al. 2011).

2.2.4 Prognostic score matching

The *prognostic score* is an alternative balancing score (Hansen 2008b). For binary outcomes, one prognostic score is the probability of experiencing the outcome, in the absence of treatment,²² conditional on the set of observed baseline variables. More generally, the prognostic score is defined as a function of the observed baseline variables such that, when conditioning on the prognostic score, outcomes in the absence of the treatment are independent of those baseline variables:

$$P(Y(0)|X, p(X)) = P(W|p(X))$$

Hansen showed that, at any value of the prognostic score, the difference between the outcomes of treated and untreated patients is an unbiased estimate of the average treatment effect at that value of the prognostic score, provided that (Hansen 2008b):

- 1) There is no unobserved confounding;
- 2) There is no level of the prognostic score at which treatment is received with certainty; and
- 3) There is no effect modification, *i.e.* no baseline variable that modifies the treatment effect.

²¹ Note that the theorems underlying propensity score matching do not require pairwise similarity of matches, but only that the baseline covariates of intervention and matched control patients come from the same multivariate distribution.

²² If there are more treated than untreated subjects, it may be appropriate to reverse the definition of 'treatment' and model outcomes under treatment rather than outcomes under the absence of treatment.

The 'no effect modification' assumption arises because outcomes in the absence of treatment are only known for untreated individuals. Therefore, one must fit a model of the outcome to untreated individuals and extrapolate this model to treated individuals.²³

Just like the propensity score, the prognostic score must be estimated from the data (often using a logistic regression model). Checking balance is important because estimated treatment effects will be biased if the model is misspecified.

2.2.5 Genetic matching

Genetic matching generalises Mahalanobis distance matching by using a more flexible multivariable distance measure (Diamond and Sekhon 2013):

$$d(X_i, X_j) = \{(X_i - X_j)(S^{-1/2})'WS^{-1/2}(X_i - X_j)\}^{1/2}$$

Here, W is a semi-definite weight matrix, the non-diagonal elements of which are taken to be zero. $S^{-1/2}$ is the Cholesky decomposition of S , the covariance matrix of X . The baseline covariates included in X are those that are instrumental to obtain adequate balance,²⁴ and these may include the estimated propensity score or prognostic score.

The genetic matching algorithm aims to optimise balance by selecting the most appropriate weight matrix. The particular measure of balance might be, for example, the minimum p-value from t-tests on the baseline variables or from Kolmogorov-Smirnov tests on these variables (this minimum p-value should then be maximised). Genetic matching has been shown to dominate propensity score matching in terms of precision in simple simulations with independent normally distributed baseline variables (Sekhon and Grieve 2012). Furthermore, a comparison between genetic matching and propensity score matching in the context of an observational study of a cardiac monitoring device showed that genetic matching resulted in better balance on empirical quantile-quantile plots of baseline variables, and also replicated the results of an RCT (Sekhon and Grieve 2012). Genetic matching also produced lower bias and greater precision than propensity score matching in simulations in which the treatment assignment mechanism differed between subgroups (Kreif et al. 2012).

2.2.6 Combined matching and regression approaches

Since matching is not an estimator in itself, some other analytical technique is needed to estimate the treatment effect once the matching has been completed (Ho et al. 2007). One simple estimator,

²³ If there is effect modification, one way to proceed is to capture the conditioning variable in the conditioning statement.

²⁴ By default, X is the same as the set of variables over which balance is assessed, but this may not always be the case. For example, the propensity score might be included in X but omitted from the final assessment of balance. Although matching on the propensity score may aid balance on important prognostic variables, balance on the propensity score might not be an aim in its own right.

described above, is the difference in the mean outcomes of the two groups. However, given concerns about misspecification of the propensity score model, recent developments in the literature have addressed methods that combine matching with regression. In general, missing-data methods have the 'doubly-robust' property if they remain consistent when either a model for the missingness mechanism (*e.g.*, propensity score) or a model for the distribution of the complete data is misspecified (as long as both are not misspecified) (Bang and Robins 2005). Examples include weighted regression, targeted maximum likelihood estimation, or bias-corrected matching (Kreif et al. 2014). One relatively simple doubly-robust approach is to fit regression models to propensity-score matched data sets.²⁵ This approach was found to reduce bias in comparison to weighted regression approaches (Kreif et al. 2013).

2.3 Design considerations in observational studies

In general, strategies to deal with threats to validity can be divided into those that occur at the design stage and those that occur at the analysis stage, where 'design' includes all contemplating, collecting, organising and analysing of data that takes place prior to seeing any outcome data, and analysis includes any part of the study that requires outcome data (Rubin 2007). Design is considered as being more critical than analysis in observational studies, because the 'selection on observables' assumption is usually a prerequisite of the analytical methods and so must be addressed by design (Rubin 2008).

All of the methods reviewed above presupposed the existence of a data set, and so did not address aspects of design that occurred before this data set was defined. This Section reviews the literature on two important design issues: the definition of the baseline variables and the population from which the matched control patients are selected.

2.3.1 Variable selection

The omission of a confounder from the propensity score model will lead to bias (Drake 1993), but less obvious historically was the effect of omitting a variable that is related to treatment, but not outcome. One case study and two sets of simulations have now shown that excluding these variables does not worsen bias in the estimated treatment effect, but can improve precision (Austin, Grootendorst, and Anderson 2007; Brookhart et al. 2006). Because achieving balance on prognostic variables is crucial, methods have been proposed to increase the priority given to the most prognostic variables in matching algorithms (Ramsahai, Grieve, and Sekhon 2011).

Austin and others considered the relative advantages of using administrative and clinical data for matching (Austin et al. 2005). They exploited observational data for people discharged alive from hospital after acute myocardial infarction between 1991 and 2001 and found that propensity score matching on administrative data was not sufficient to balance characteristics recorded in clinical data.

²⁵ Some weak conditions are required for this method to have the doubly-robust property, *e.g.*, to ensure that poor overlap does not mean that some cases are not matched, even though the correct parametric model is specified (Ho et al. 2007).

While the availability of large administrative data sets may make propensity score methods seem attractive, efforts must be placed on collecting the most appropriate data sets, which may include clinical data (Stuart et al. 2013). In many instances, however, the set of variables is constrained by the available data, and efforts must either focus on assessing the impact of residual unobserved confounding (for example, using simulations) or on careful selection of the population of control patients.

2.3.2 Source of control patients

One design issue that has received scant attention relates to the choice of higher-level unit from which to select the control group. This issue arises because often an intervention is piloted within a sample of units (*e.g.*, hospitals, geographic areas or schools). Within these units, a subset of individuals will receive the intervention and others will receive the control. However, data are often available for other potential control units. Thus, there are several possible strategies for selecting matched control patients:

1. From within the intervention area ('local controls');
2. From all areas (*e.g.*, national sample);
3. From a convenience sample of areas; or
4. From areas selected purposively.

Several approaches have been used in the applied literature. For example, an evaluation of strategies for eliminating substance misuse for selected Medicaid enrollees in Oregon selected control patients from within the Oregon area (McConnell et al. 2008), while some Medicare Disease Management and Care Coordination Demonstrations were evaluated by matching patients to similar patients from other providers (Nelson 2012), and several pilots of coordinated care (Roland et al. 2012) have been evaluated against control patients selected nationally.

The choice of control population clearly has the potential to affect the extent of confounding through area-level variables, because these variables would be balanced by design when using local controls but not when using other sources of controls. Less obvious are the effects on confounding through individual-level variables. To explore the issues, suppose that, within the intervention area, a particular characteristic (*e.g.*, female gender) is associated with treatment receipt. Thus, the local untreated group will contain relatively few women. By contrast, in those areas not offering the intervention, there may be a relatively high proportion of women. Hence, if controls are selected from area(s) other than the intervention area, it might be possible to improve balance on gender relative to local controls. The same argument applies to unobserved variables, for which balance may be improved or made worse by selecting controls from outside of the intervention area. Therefore, the choice of control area can be seen as a trade-off between the risks of confounding at the individual versus area levels.

Defining control populations has long been recognised as a central issue for case-control studies (Miettinen 1985).²⁶ For cohort studies, methodological attention to this issue has been haphazard. Some concern arose indirectly from LaLonde's work in the 1980s (LaLonde 1986). LaLonde was concerned about the internal validity of observational studies, and he re-analysed data collected for an RCT of a job training programme, namely the National Supported Work Demonstration that operated in 10 U.S. States. In his work, he compared the outcomes of the RCT intervention group with people from two national surveys, using difference-in-difference and Heckman selection models (Heckman 1979). The effect sizes thus obtained differed from the experimental benchmark obtained from the RCT data alone, leading him to conclude that policymakers should be wary of results from observational studies. As LaLonde selected observational controls from two national surveys, presumably the majority of his controls came from U.S. States other than the 10 that participated in the intervention.

LaLonde's study was followed by a number of other studies that compared RCT and observational study estimates for similar interventions (these are known as 'within study comparisons').²⁷ One review of these found that randomised benchmarks are more likely to be recovered if controls are selected from within the intervention area (Cook, Shadish, and Wong 2008). Another review, which focussed exclusively on job training and employment interventions, concluded that, in multivariable analysis, discrepancies from RCT benchmarks were increased when observational controls were selected from matched geographic areas rather than from within the intervention area, but not by as much as when selected nationally or from areas apparently selected as part of a convenience sample (Glazerman, Levy, and Myers 2003). On the basis of these within-study comparisons, it seems reasonable to suppose that a control area strategy that selects controls from the same area as the intervention results in treatment effect estimates that are closer to the RCT benchmark.

Unfortunately, reviews of within-study comparisons are unlikely to be tell us much about which control area strategies provide the least biased, most efficient estimates of treatment effect (Deeks et al. 2003).²⁸

Some problems with within-study comparisons are that the number of constituent studies has been limited and dominated by labour market interventions. More fundamentally, although RCTs are often

²⁶ In case-control studies, groups are defined based on their outcomes rather than, as in the current context, their exposure to the treatment of interest (Rothman, Greenland, and Lash 2008).

²⁷ Cook and colleagues defined within-study comparisons as 'studies that take the effect size from an experiment and contrast it with the effect size from an observational study sharing the same treatment group' (Cook, Shadish, and Wong 2008). These have not all relied on analytical methods such as regression adjustment to adjust for differences in the characteristics of patients across RCT and observational settings. For example, in the trial by Shadish and colleagues, undergraduate students were randomly assigned to an RCT of vocabulary training versus mathematics training, or to a prospective observational study of the same two trainings (Shadish, Clark, and Steiner 2008).

²⁸ The same arguments apply to 'meta-epidemiological' (or 'between study') approaches that aim to identify published studies of similar interventions and compare treatment effects according to aspects of the methods used (Deeks et al. 2003).

considered to be the gold standard in comparative effectiveness research, in some cases there are legitimate reasons why observational studies should give different treatment effects to RCTs, for example because of Hawthorne effects (McCarney et al. 2007). The constituent studies may differ in many more ways than the source of control population, and these differences (such as the treatments, populations of interest, or estimation procedures) often cannot be fully controlled for in meta-epidemiological work (Shadish, Clark, and Steiner 2008). These observations may explain why findings from within-study comparisons have not been entirely consistent. Dehejia and Wahba (2002), for example, re-analysed LaLonde's data with propensity score matching and were able to replicate the experimental benchmark, even though controls were selected from national panel data.

A few studies have sought to examine issues about control area selection directly. Although he did not address questions about which of several control areas should be used, Rosenbaum (1987) explored how multiple control groups could be used to test for unobserved confounding due to individual-level variables. He argued that, if outcomes for individuals in the two control areas are different after adjustment for observed variables, then there must be unobserved confounding. His approach did not explicitly reflect area-level confounding. Multiple control groups are sometimes used in the medical literature, for example by comparing treated patients to both historic and concurrent controls (Harrison et al. 2010), but they seem to be rare in comparison to studies that use a single group (Austin 2008a).

Stuart and Rubin (2008) considered a situation in which:

- Data were available from two higher-level units (the one that offered the intervention and one other);
- There was limited overlap between the baseline characteristics of intervention and control patients within the intervention unit; and
- There was no unobserved confounding.

For this situation, Stuart and Rubin derived approximate equations for the number of matched controls that should be sourced from within the intervention unit rather than from the other unit, so as to minimise bias.²⁹ Further, they proposed an approach whereby the intervention effect thus estimated could be corrected for the bias that is introduced through confounding from area-level variables. They did not consider the scenario in which there was unobserved confounding at the individual level. Unfortunately this is a crucial factor for observational studies, which might have differential effects on bias according to the choice of control population. In summary, issues about control area selection have not been sufficiently addressed.

²⁹ They assumed that all intervention patients should be matched.

2.4 The generalisability of RCTs

The main advantage of a well-executed RCT is that it can balance observed and unobserved variables between treatment arms. However, inferences from RCTs might not apply to the target populations of policy relevance. Efforts have been made at the design stage to maximise the generalisability of RCTs, such as through the use of pragmatic trial designs (Schwartz and Lellouch 1967; Roland and Torgerson 1998). Unlike explanatory trials that compare highly-specified interventions in controlled situations on tightly defined cohorts of patients, pragmatic trials use broad eligibility criteria to enrol patients, test flexible interventions as practiced by typical (not expert) practitioners, and make comparisons with usual care (DeMets 2012). Despite their advantages, pragmatic trials can still have problems with generalisability. For example, they sometimes select sites purposively rather than randomly (Olsen et al. 2013). Furthermore, pragmatic trials might not reflect the patterns of service delivery that exist in routine settings (Basu 2012), for example because the trial protocol might constrain the development of the services (Hendy et al. 2012).

Comparatively little attention has been placed on quantitative, analytical-stage strategies to assess and extend the generalisability of RCTs. For example, the Consolidated Standards of Reporting Trials (CONSORT) statement, which was first published in 1996, requires only a “table showing baseline demographic and clinical characteristics for each group” (Schulz, Altman, and Moher 2010). Stuart and colleagues examined issues about generalisability more directly, and extended CONSORT-style guidelines (Stuart et al. 2011) in two ways. Specifically:

- They proposed using the propensity score as a measure of the similarity of RCT participants and people in the general population (their propensity score related to the probability of being in the trial, and they proposed comparing the average scores of the two groups); and
- They used the propensity score to reweight outcomes for the RCT control group to reflect the characteristics of the general population.

In their case study of a school-wide behaviour improvement programme, the reweighted outcomes of the trial control schools tracked those of the wider population of local schools. A similar method was applied to reweight the treatment effect of a pharmacological intervention for Human Immunodeficiency Virus (HIV) estimated using RCT data to the population of people infected with HIV in the United States (Cole and Stuart 2010).

Hartman and colleagues were also concerned with methods to reweight trial estimates to the general population (Hartman et al. 2015). Their approach involved reweighting the estimates of SATT produced from the RCT according to the characteristics of people receiving the intervention in routine practice to obtain an estimate of PATT.³⁰ An important intermediate step was to reweight the trial intervention group outcomes to the characteristics of people receiving the intervention in routine

³⁰ This differs from work of Cole and Stuart, who estimated PATE.

practice, and to test for similarity in the adjusted outcomes of these groups using placebo tests. These placebo tests aimed to test the validity of some of the assumptions made when estimating PATT. Because the two groups were assumed to receive the same treatments, Hartman and colleagues argued that marked differences could only plausibly result from bias due to the reweighting method. Unlike standard tests, which aim to establish whether there is sufficient evidence of difference, these placebo tests assess the evidence for similarity (Jones 2007; Hartman, Grieve, and Sekhon 2010). Hartman and colleagues adopted as their null hypothesis that the RCT and population treated groups had meaningfully different outcomes, with the alternative hypothesis being that their outcomes were sufficiently similar. This put the onus on the investigators to demonstrate that, after reweighting the RCT data, the outcomes for the intervention arm of the RCT were sufficiently similar to those of people receiving the corresponding treatment in the target population. Hartman and colleagues applied placebo tests to a study of a surgical intervention, namely pulmonary artery catheterisation (Hartman et al. 2015).

2.5 Discussion

Issues about the internal validity of observational studies and the generalisability of RCTs are closely related, since both are fundamentally a problem of selection bias. For RCTs, the assignment mechanism concerns recruitment into the sample, rather than selection into the intervention in routine practice, as in observational studies. However, in both cases, the assignment mechanism is usually unknown and so must be estimated from the data.³¹ The framework introduced at the beginning of this Chapter makes clear that both selection mechanisms must be addressed in order to produce unbiased inferences of population average treatment effects (Imai, King, and Stuart 2008).

The discussion that follows is motivated by the observation that most of the methodological work to address the internal validity of observational studies has relied on innovations at the analysis stage,³² while most of the methodological work to address the generalisability of RCTs has focussed on the design stage. Thus, there is scope to improve the design of observational studies by applying methods developed for RCTs, and likewise to improve the analysis of RCTs by applying methods developed for observational studies.

2.5.1 *Gaps in the literature regarding the internal validity of observational studies*

It is clear that many analytical methods have been developed to deal with internal validity in observational studies, and these are being vigorously compared and evaluated (Kreif et al. 2012; Stürmer et al. 2006; Stukel et al. 2007). In contrast, much less attention has been placed on choosing the geographic area (or, more generally, higher-level unit) from which controls are selected. These

³¹ Except for in the comparatively rare instances when individuals from the general population can be randomised into the trial - ethical considerations usually do not permit this.

³² Strictly, according to the framework proposed by Rubin (2007), methods such as propensity score matching are design-stage methods, as they do not use the outcome data. However, the observation remains that many advances in observational methods presuppose the existence of a data set.

issues are complex as they involve trade-offs between confounding at the individual versus area levels, and they are likely to become more relevant as large, routine data sets become more available for research (Pope et al. 2014).

None of the studies described in Section 2.3.2 assessed which strategy for control area selection minimises bias and mean-squared error and under which circumstances. The most comprehensive study (Stuart and Rubin 2008) assumed no unobserved confounding at the individual level, which is unrealistic (Rubin 2010). To address this gap in the literature, this thesis includes a case study and a set of simulations to compare four strategies to control area selection: local controls, controls from a matched area, controls from a sample of areas, and national controls (Chapter 4).

2.5.2 Gaps in the literature about the external validity of RCTs

An emerging body of work is exploring how observational methods can be used to assess and extend the generalisability of RCTs (Stuart et al. 2011; Cole and Stuart 2010; Hartman et al. 2015; Hartman, Grieve, and Sekhon 2010). These methods are significant steps forward on previous approaches that relied on interpreting the 'representativeness' of patients from tables of baseline characteristics (Schulz, Altman, and Moher 2010) or on making other comparisons between outcomes from randomised and nonrandomised settings (Cook, Shadish, and Wong 2008).

It is informative to note parallels between this work and within-study comparisons such as LaLonde (1986).³³ Although LaLonde and others were chiefly concerned with the internal validity of observational studies, their analyses can instead be seen as assessments of the generalisability of the RCTs. Indeed, their focus on internal validity is problematic because the treatment assignment mechanism that they examined related to selection into an RCT rather than, as in most observational studies, treatment selection in routine practice. Because selection into an RCT is fundamentally a problem of generalisability, many of the lessons that emerged from within-study comparisons are relevant to this newer work about generalisability. For example, sets of quality criteria for within-study comparisons have generally highlighted the need to (Cook, Shadish, and Wong 2008):

1. Use objective measures of what constitutes 'similarity' in outcomes;
2. Conduct nonrandomised adjustments according to best practice in this area;
3. Ensure that the groups receive the same treatments; and
4. Standardise measurement across groups.

³³ It is also noteworthy that many of the methods developed by Stuart and others are similar to those that have been applied to answer questions about the effect of RCT protocols on those recruited. For example, Gross and colleagues noted that there is a perception that patients are likely to have better outcomes when treatment decisions are based on physician's clinical judgement, rather than random assignment as in an RCT (Gross et al. 2006). Other studies have assessed whether participation in an RCT may improve outcomes regardless of allocation, for example because of the Hawthorne effect (McCamey et al. 2007; Marcus et al. 2012).

The placebo tests used by Hartman and colleagues can be seen as an attempt to use objective measures (criterion 1 above) and furthermore, to conduct nonrandomised analysis in the best way possible (criterion 2). However, in applying placebo tests to a narrowly defined, surgical intervention, Hartman and colleagues did not specifically address criterion 3. This aspect is particularly important when evaluating complex interventions, since these are context dependent and so there is a greater risk that the interventions tested within an RCT setting are not the same as those delivered in routine practice. A practical consequence of these observations is that, when applying placebo tests to complex interventions, it is important to know how one should proceed when the placebo tests fail. This is an important topic for policymakers because, if a placebo test fails to confirm the generalisability of an RCT, then it may be unclear whether the intervention is effective in the target population.

This thesis extends the placebo tests of Hartman and colleagues in the following ways:

1. The strategy used to select controls for RCT participants is informed by the prior study about control area selection (Chapter 4);
2. Variable selection for the placebo tests is informed by a qualitative study that examined the reasons given by patients to refuse to participate in the RCT (Sanders et al. 2012);
3. The placebo tests exploit large, linked routine data sets, and in so doing represent a general category of trials that are embedded within these data sets. The nature of the data sets have some implications for the estimation procedure used by the placebo tests, with a focus on time series methods; and
4. I propose a new form of sensitivity analysis for when the placebo tests fail, in order to examine the implications of alternative causes of non-generalisability.

2.6 Structure of the remainder of this thesis

Chapter 3 presents the findings from the observational case study, which concerns telephone health coaching. Chapter 4 then examines a methodological issue that emerged from the case study as well as from the above discussion, namely the optimal source of control population. Chapter 5 turns to RCTs, and presents findings from the Whole Systems Demonstrator (WSD) trial of telehealth. The WSD embodies many of the challenges to generalisability faced when evaluating complex interventions within RCTs. These challenges are discussed in more detail in Chapter 6, which develops the placebo tests and sensitivity analyses. Chapter 7 draws on the experiences of each of these studies, plus more general theory about complex interventions, to propose a general approach to evaluation in this area, while Chapter 8 concludes the thesis.

Chapter 3 Effect of telephone health coaching (Birmingham OwnHealth) on hospital use and associated costs: cohort study with matched controls (research paper 1)

3.1 Preamble to research paper 1

The first paper is an observational analysis of an established telephone health coaching service. Over 3,000 health-coached patients were linked to national hospital administrative data, and the majority of these were compared with matched controls. Controls were selected from areas of England that were first matched to the characteristics of the intervention area. The person-level matching algorithm was based on prognostic scores, which were thought to be particularly suitable for analyses of the effect of complex interventions on hospital admissions (Box 2). Following enrolment, rates of emergency hospital admission were found to increase more quickly among health-coached patients than matched controls.

The observational methods used in this paper should aid generalisability. For example, the study was retrospective, so could not have influenced the care received by patients during the study period. Furthermore, the study concerned an intervention that was delivered as part of routine health care to a large number of patients. However, as discussed in Chapter 2, observational methods face particular threats to internal validity. Therefore, this study included sensitivity analysis based on simulating the implications of hypothetical unobserved confounders for the estimated treatment effect. Thus, the analysis quantified the strength of unobserved confounding that would be required to reverse the findings of the study for emergency hospital admissions. The required associations were larger than those suggested from the prior literature, so it seems unlikely that unobserved confounding obscured reductions in admissions in this instance.

While the study has strengths, it raises questions about how best to deal with the possibility of unobserved confounding. Chapter 4 addresses an important aspect of study design that might influence the size of the bias due to unobserved confounding, namely the choice of population from which the matched controls are selected.

This paper has previously been published in the *BMJ* (Steventon et al. 2013b), under the Creative Commons Attribution Non Commercial (CC BY-NC 3.0) license. It has been reformatted for this thesis.

Box 2: The advantages of prognostic scores in evaluations of complex interventions

Prognostic score matching offers potential advantages over propensity score matching when evaluating the effect of complex interventions on hospital admissions. Many studies have examined the relationships between hospital admissions and both patient and organisational characteristics (Ross et al. 2008; Tsai et al. 2013; Billings et al. 2006; Bardsley et al. 2011; Billings et al. 2013). In contrast, the relationship between treatment assignment and patient characteristics appears to be less well understood (Mauri 2012). This difference in knowledge might have arisen because hospital admissions are recorded routinely across large populations (Spencer and Davies 2012), and so can be relatively easily studied. In contrast, treatment allocation decisions in routine practice are the product of private discussions between patients and health care professionals, and it is not always clear which alternative treatment options were considered.

Prognostic score matching requires different assumptions to propensity score matching (Chapter 2). Ultimately, the choice of matching method should be based on achievement of adequate balance (Ho et al. 2007), something that was obtained after prognostic score matching in this study.

RESEARCH PAPER COVER SHEET

PLEASE NOTE THAT A COVER SHEET MUST BE COMPLETED FOR EACH RESEARCH PAPER INCLUDED IN A THESIS.

SECTION A – Student Details

Student	Adam Mark Charles Steventon
Principal Supervisor	Professor Richard Grieve
Thesis Title	Evaluating complex interventions using routinely collected data: Methods to improve the validity of randomised controlled trials and observational studies

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?	BMJ		
When was the work published?	2013		
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?*	No	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	Choose an item.

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 contribution included the overall design of the study, including the approach to matching, estimation and sensitivity analysis. I liaised with Birmingham East and North Primary Care Trust to arrange for the transfer of identifiable data to the NHS Information Centre for health and social care for data linkage. I conducted the analyses and produced the first draft of the manuscript, before incorporating comments from other co-authors and from peer reviewers.
--	--

Student Signature: _____

Date: _____

Supervisor Signature: _____

Date: _____

3.2 Abstract

Objectives: To test the effect of a telephone health coaching service (Birmingham OwnHealth) on hospital use and associated costs.

Design: Analysis of person-level administrative data. Difference-in-difference analysis was done relative to matched controls.

Setting: Community-based intervention operating in a large English city with industry.

Participants: 2,698 patients recruited from local general practices before 2009 with heart failure, coronary heart disease, diabetes, or chronic obstructive pulmonary disease; and a history of inpatient or outpatient hospital use. These individuals were matched on a 1:1 basis to control patients from similar areas of England with respect to demographics, diagnoses of health conditions, previous hospital use, and a predictive risk score.

Intervention: Telephone health coaching involved a personalised care plan and a series of outbound calls usually scheduled monthly. Median length of time enrolled on the service was 25.5 months. Control participants received usual health care in their areas, which did not include telephone health coaching.

Main outcome measures: Number of emergency hospital admissions per head over 12 months after enrolment. Secondary metrics calculated over 12 months were: hospital bed days, elective hospital admissions, outpatient attendances, and secondary care costs.

Results: In relation to diagnoses of health conditions and other baseline variables, matched controls and intervention patients were similar before the date of enrolment. After this point, emergency admissions increased more quickly among intervention participants than matched controls (difference 0.05 admissions per head, 95% confidence interval 0.00 to 0.09, $p=0.046$). Outpatient attendances also increased more quickly in the intervention group (difference 0.37 attendances per head, 0.16 to 0.58, $p<0.001$), as did secondary care costs (difference £175 per head, £22 to £328, $p=0.025$). Checks showed that we were unlikely to have missed reductions in emergency admissions because of unobserved differences between intervention and matched control groups.

Conclusions: The Birmingham OwnHealth telephone health coaching intervention did not lead to the expected reductions in hospital admissions or secondary care costs over 12 months, and could have led to increases

3.3 Introduction

Facing rising costs, health care systems around the world are exploring innovative ways to improve efficiency. Particular attention has been placed on the use of technology to help manage long-term health conditions (McLean, Protti, and Sheikh 2011), including one-to-one telephone health coaching. This involves a regular series of phone calls between patient and health professional. The calls aim to provide support and encouragement to the patient and promote healthy behaviours such as treatment control, healthy diet, physical activity and mobility, rehabilitation, and good mental health (Hutchison and Breckon 2011). The hope is that the patient will maintain their own health more independently, and that the professional and patient will be in a better position to identify problems before they become critical. In turn, admissions to hospital may be prevented (Birmingham OwnHealth 2007). Avoidable hospital admissions are both undesirable for the patient and expensive for the payer.

In a systematic review of telephone health coaching for people with long-term conditions, only nine of 34 studies investigated effects on health service use (Hutchison and Breckon 2011). Four studies showed effects in this area, but findings were hard to generalise because the studies looked at a range of different health conditions, had relatively small samples (the average sample was fewer than 360 people), and included interventions that were heterogeneous. Five of the nine interventions included telemonitoring of vital signs such as blood pressure alongside telephone coaching. Since the review, other larger studies have been conducted.

Wennberg and colleagues conducted a large randomised controlled trial of 174,120 patients with employer-based health insurance (Wennberg et al. 2010). The intervention was given to patients who were at high predictive risk (for example, of future hospital costs), with a lower risk threshold used in one treatment group than in the other. The researchers concluded that telephone care management reduced hospital admissions overall and among patients with selected long-term conditions (heart failure, coronary artery disease, Chronic Obstructive Pulmonary Disease (COPD), diabetes, and asthma), although there was no statistically significant effect on the subset of admissions that came through the emergency room. Among the subgroup of patients with long-term conditions, overall medical and pharmacy costs were \$51 (£33.8; €38.8) per month lower in the more aggressively targeted group than the other. Their intervention did not include telemonitoring, but it did include an element of shared decision making for 'preference-sensitive conditions' (for example, with regards to treatment options for arthritis of the hip). The shared decision-making could be largely responsible for the intervention's effect on admissions that did not come through the emergency room (Motheral 2011).

Further, Lin and colleagues studied telephone health coaching for 874 US Medicaid members of working age (range 18-64 years, mean 45.3) with at least one of 10 qualifying long-term conditions and at least two acute hospital admissions or emergency department visits within a 12-month period

(Lin et al. 2012).³⁴ Their intervention relied solely on phone calls and mailed educational materials. Health coaches provided information on conditions and treatment options, empowered patients to self-manage and self monitor their conditions, and encouraged patients to communicate their preferences to providers. Compared with a matched control group, the authors found no effects on hospital use and expenditures over one year. However, over two years, the number of emergency department visits reduced by a smaller amount for intervention patients than for matched controls, leading to a relative 20% increase in emergency department visits for intervention patients. Findings over one year were later reproduced in a randomised controlled trial among Medicaid patients (Kim et al. 2013).

There is continued interest in telephone health coaching with several providers active in this area, but the evidence base is unclear. As only a limited number of large studies have examined the effects on hospital use and have produced contradictory findings (Wennberg et al. 2010; Lin et al. 2012), more information is needed to understand the elements that make up a successful service. We were commissioned by the Department of Health in 2010 to evaluate the effect of England's largest example of telephone health coaching (Birmingham OwnHealth) on hospital use and associated costs. Previous evaluations of Birmingham OwnHealth had shown that patients had high levels of satisfaction and believed that the service reduced their need to go to hospital (Azarmina and Lewis 2007). In addition, a study showed reductions in levels of glycated haemoglobin (HbA1c), blood pressure, and body mass index among a subset of patients with poorly controlled diabetes (Jordan, Lancashire, and Adab 2011). We present estimates of the effect of the Birmingham OwnHealth service on hospital admissions and associated costs. The service was decommissioned in 2012 after a consultation exercise (NHS Direct 2011).

3.4 Methods

In summary, a retrospective design was used to assess the effects of an existing service. Matched controls aimed to reflect the changes in hospital use that can occur over time even without an intervention (Steventon et al. 2012b; Roland et al. 2005).

3.4.1 Intervention, including patient recruitment

Birmingham OwnHealth was established in 2006 in a large city with industry. This area had health inequalities and some parts with high deprivation. Birmingham OwnHealth aimed to improve self-care strategies, improve clinical indicators, and reduce health service use (Birmingham OwnHealth 2007).

³⁴ Medicaid is a publicly funded health insurance programme for families and individuals with low income and limited resources in the United States.

The service targeted people with heart failure, coronary heart disease, diabetes, or COPD. Inclusion criteria were all of the following:

- A recorded diagnosis of one of the targeted conditions;
- A minimum level of disease severity (for example, HbA1c >7.4 in the past 15 months) (Birmingham OwnHealth 2007);
- Age 18 or older;
- Ability to communicate on the telephone; and
- A recorded address and practice registration.

Potentially eligible patients were identified through analysis of data extracts sourced from the participating general practices. Summary files were then reviewed and ratified by general practitioners. General practitioners applied additional clinical judgment in determining which patients to refer into the Birmingham OwnHealth service, based on a set of consideration factors. These factors included comorbidities under active treatment, personality and mental health problems, and life circumstances such as pregnancy. General practice staff sent an introductory letter to the selected patients, which was followed by a phone call by a representative from Birmingham OwnHealth.

Once enrolled, patients (who were known as 'members') were assigned a care manager, who were specially trained nurses employed by NHS Direct. General practice data were also transferred onto the operational systems used by Birmingham OwnHealth. During the programme, care managers followed five fundamental steps: assessment, recommendation, follow-up, ongoing management, and review.

Care managers made regular telephone calls to patients. These calls were usually made monthly at a predetermined date and time to suit the user, although a minority of patients received calls more frequently (two to four times per month) owing to disease severity, social isolation, or severe weather. As the number of members grew, some members were stepped down (or 'graduated') to quarterly calls. During the telephone calls, care managers asked patients about current health status and symptoms, and recorded this information along with other information such as recent test results and changes to treatment. Care managers then gave personalised guidance and support, aiming to build continuing relationships with patients and provide motivation, skills, and knowledge to encourage patients to better manage their health conditions. The calls focussed on eight priorities in care management—namely, to ensure that patients were able to do all of the following:

- Know how and when to get help for health and social care problems;
- Learn about their condition, and agree and set treatment goals within a personalised care plan;
- Take medicines correctly;

- Get recommended tests and services;
- Act to keep the condition in good control;
- Learn how to make changes to lifestyle and circumstances to reduce risks;
- Build on strengths and overcome obstacles, while strengthening personal social networks; and
- Follow up with specialists and appointments.

Calls were structured into modules focussing on each of these priorities, and used screen-prompted algorithms to structure the conversation. The software prompted care managers to follow guidelines in each priority area (for example, provide basic information about treatment). Care managers, however, were not constrained to follow protocols. Additional educational materials could be sent to patients.

Care managers aimed to coordinate input across services, for example, when developing personalised care plans, and could refer patients onto existing services such as mental health services and social care. General practitioners were offered monthly phone calls and quarterly meetings with their assigned care manager to discuss patients. The service was set up to provide proactive calls to patients rather than act as a phone-in service, and inbound calls were less common, comprising about 5% of calls.

3.4.2 Study populations

Study participants were enrolled in Birmingham OwnHealth between the time the intervention commenced in April 2006 and December 2008. This cut-off period was chosen to ensure sufficient time to follow-up for at least 12 months when this study was commissioned. The service provider identified all such patients recorded in their operational datasets, from which we excluded those without a record of inpatient or outpatient hospital use in the three years before enrolment. Because the matching variables came from hospital data, we could not accurately characterise patients without previous hospital activity. However, we would expect these excluded patients to have low future levels of hospital use (Steventon et al. 2012b), and therefore represent little potential to reduce hospital costs over 12 months.

The large size of the service reduced the scope to find matched controls locally, and there may have been spillover effects. Therefore, we selected several comparable areas within England to provide a pool of potential matched controls. Four of these areas (Bradford and Airedale, Sandwell, Stoke, and Wolverhampton) were drawn from a national area classification (Office for National Statistics 2010), which were similar to the intervention area in terms of demography; occupational mix; and rates of education, occupation, limiting long term illness, and unpaid care. We also included Walsall, which was commonly used as a comparator by the Primary Care Trust that commissioned Birmingham OwnHealth. We checked that a similar health coaching service via telephone had not operated in the

selected areas, using Internet searches and discussions with colleagues. As a result, we excluded parts of Walsall from the pool.

3.4.3 Study endpoints and sample size calculation

Our primary endpoint was the number of urgent and unplanned ('emergency') hospital admissions per head over 12 months. Our hypothesis was that the service could alter rates of emergency admissions in either direction. Increases in emergency admissions have been suggested by studies of other types of interventions involving patient outreach in England (Gravelle et al. 2007).

We performed a sample size calculation at the outset of the study to check that we were likely to have data for a sufficient number of patients to produce meaningful conclusions. We thought it important to detect relative changes of 15% should they occur, based on the level of effect judged as meaningful in similar studies (Steventon et al. 2012a), at 90% power and with two-sided $p=0.05$.

Annual admission rates for the usual care group were assumed to be 0.25 per person with a standard deviation of 0.4 (The Health & Social Care Information Centre 2013). Calculations were performed in SAS 9.3, and assumed a correlation of 0.15 between the number of admissions for intervention and matched control patients. Based on these assumptions, 2,035 intervention patients were needed.

Secondary endpoints calculated over 12 months included the number of planned ('elective') hospital admissions, number of hospital bed days, number of attendances for hospital-based ambulatory care ('outpatient attendances'), and costs of secondary care.

3.4.4 Data sources and data linkage

The service providers had access to identifiable data for participants, including a national patient identifier (the 'NHS number'), sex, date of birth, and postcode. These data were transferred to the NHS Information Centre for health and social care who used them to link participants to national administrative data for secondary care activity (the Hospital Episode Statistics, or HES) (The Health & Social Care Information Centre 2013). The data linkage required an exact or partial match on several of the variables at once. After the data had been linked, the HES identity was transferred to the evaluation team together with the date of patient enrolment into Birmingham OwnHealth, year of birth, sex, and small geographic area code. As a result, we only had access to 'pseudonymised' data in which all identifiable fields had been removed or encrypted. The ethics and confidentiality committee of the National Information Governance Board confirmed that data could be linked in this way without explicit patient consent.

3.4.5 Variable definitions

For intervention patients, study endpoints were calculated over 12 months after the date of enrolment into Birmingham OwnHealth. For matched controls, we used the date of enrolment for the corresponding intervention patient. Variables were therefore calculated over the same period for matched pairs as for intervention patients.

Analysis of inpatient activity was limited to 'ordinary admissions' by excluding regular ward attendances, maternity events, and transfers. Admissions were then classified into either emergency or elective admissions, based on the method of admission. Bed days included stays after emergency admission only and excluded same day admissions and discharges.

Secondary care costs included inpatient and outpatient costs. They were estimated by applying a set of unit costs specific to the case mix (Department of Health 2007), which represented the tariff amounts that providers were allowed to charge commissioners in the United Kingdom's health service. We did not include adjustments for regional differences in care costs, to allow robust comparison of the volume of care services between intervention patients and matched controls. We only attached costs to activity covered by mandatory tariffs, which excluded locally negotiated non-tariff payments and augmented care payments associated with critical care.

Baseline variables were derived using hospital data recorded before the enrolment dates. The variables were based on those used in an established predictive model for emergency hospital admissions over 12 months (Billings et al. 2006). These variables were age band; sex; area-based socioeconomic deprivation score (Communities and Local Government 2008); health conditions; the number of long-term health conditions; and previous emergency, elective, and outpatient hospital use. There were 16 health condition variables, formed from primary and secondary codes from ICD-10 (International Classification of Diseases, 10th revision) on inpatient data over three years. These health conditions were anaemia, angina, asthma, atrial fibrillation and flutter, cancer, cerebrovascular disease, congestive heart failure, COPD, diabetes, history of falls, history of injury, hypertension, ischaemic heart disease, kidney failure, mental health conditions, and peripheral vascular disease.

In addition to these baseline variables, we estimated the risk of emergency hospital admission in the subsequent 12 months. The predictive risk models were based on the variables used in the published model (Billings et al. 2006) but reweighted to reflect patterns of hospital use of Birmingham residents who had never been enrolled into the service. Models were constructed using logistic regression on a monthly basis through the enrolment period and validated on split samples. The estimated β coefficients from the validated models were then applied to intervention patients and potential controls to produce the risk scores that would be used in the matching process.

3.4.6 Methods to select control group

There are several methods for selecting matched control groups, but the aim is always to select, from the wider population of potential controls, a subgroup of patients that is similar to the intervention group with respect to predictive baseline variables (Rosenbaum and Rubin 1983). The risk score was strongly predictive, so we used a calliper approach whereby the pool of potential matches for a given intervention patient was narrowed down to those patients with a similar risk score (within 20% of one standard deviation) (Hansen 2008b). From within this restricted set, one control was selected for

the intervention patient based on the individual baseline variables using the Mahalanobis multivariate distance metric (Rosenbaum and Rubin 1985). One matched control was selected for each intervention patient. This was done without replacement, so that the control group consisted of distinct individuals.

The main diagnostic in matched control studies is balance, which refers to the similarity of the distribution of baseline variables between intervention and matched control groups. Formal statistical tests are not recommended in the assessment of balance, because they depend on the size of the groups as well as their similarity (Imai, King, and Stuart 2008). Instead, we assessed balance using the standardised difference, which is the difference in means as a proportion of the pooled standard deviation (Austin 2008a). Although the standardised difference would ideally be minimised without limit, 10% is often used as a threshold to denote meaningful imbalance (Normand et al. 2001). The ultimate aim was to select a matched control group that was well balanced across all of the baseline variables, including the predictive risk score; therefore, we adapted the set of variables included in the Mahalanobis distance until we achieved the satisfactory balance (Ho et al. 2007).

3.4.7 Statistical approach

After matched groups had been constructed, we estimated the intervention effect using a difference-in-difference estimator. Thus, intervention and matched control groups were compared in terms of the change in the number of hospital admissions observed from the year before the enrolment date to the year after the enrolment date. Paired t-tests were conducted on the change scores to reflect the matched nature of the data (Austin 2008b).

Analysis was conducted over 12 months, regardless of death. The use of national administrative data to define variables meant that we considered there was a limited amount of missing data, because patients could be tracked even if they moved out of the Birmingham area, provided that they remained within England. We did not analyse patients who could not be linked to hospital data or could not be matched to a control.

3.4.8 Efforts to avoid bias and sensitivity analysis

The analysis was designed to reduce the susceptibility of the study to differences between intervention and control groups. Differences in variables that are predictive of future hospital use could result in confounding and biased estimates. The matching algorithm removed differences in important predictive variables and the difference-in-difference estimator was expected to remove the effect of all confounders that are not time varying, regardless of whether or not they were observed.

We conducted sensitivity analyses to test the robustness of our findings to time-varying, unobserved confounding (Groenwold, Hak, and Hoes 2009). Firstly, we followed the recommendation of West and colleagues (West et al. 2008) and compared the intervention and matched control groups in terms of an outcome that we did not expect to be influenced by the intervention—namely, in-

hospital mortality over 12 months (although one randomised study has reported effects of telephone health coaching on mortality (Alkema et al. 2007)). Secondly, we assessed the strength of unobserved confounding that would be required to alter our findings in relation to a dichotomised version of the our primary endpoint. Specifically, we simulated a hypothetical unobserved confounder and estimated the odds ratios that would be required between this confounder and intervention status and outcome for our findings to be altered (Higashi et al. 2005). The values thus obtained were compared with estimates of odds ratios for unobserved confounders, based on another study (Jordan, Lancashire, and Adab 2011).

3.4.9 Ethics approval

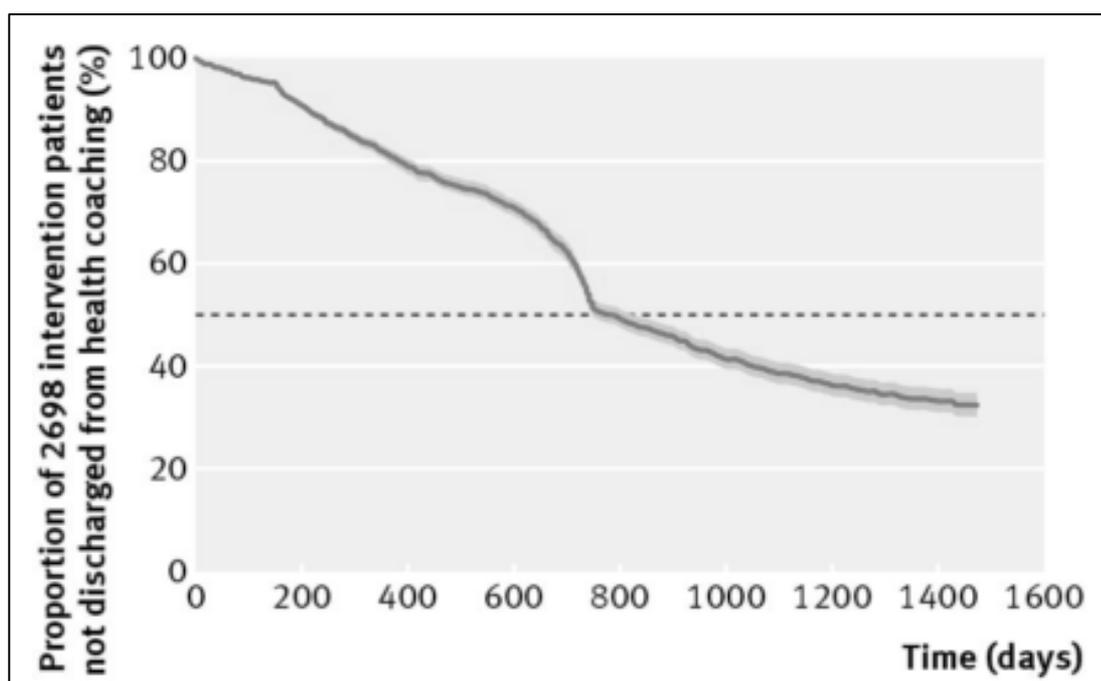
The National Research Ethics Service confirmed that ethical approval was not required for this work, because it involved retrospective analysis of non-identifiable data for the purposes of service evaluation.

3.5 Results

3.5.1 Study populations

Of 3,525 patients enrolled during the study period, 3,070 (87.1%) were linked uniquely to one individual in the hospital data. Of the 455 records that did not link to hospital data, 86% had missing or incomplete personal linkage data in the service's operational system. Of 3,070 patients linked to hospital data, 2,703 (88.0%) patients had a history of inpatient or outpatient hospital use. Matched controls were found for 2,698 (99.8%) of these patients. Data from the Birmingham OwnHealth service's operational system showed that the median duration of enrolment for the included intervention patients was 776 days (25.5 months; Figure 2). Telephone calls typically lasted for 15 minutes.

Figure 2: Numbers of days spent enrolled in telephone health coaching



Note: Solid line=best estimate, shaded area=95% confidence interval.

Predictive risk models were fitted for each of the 33 months spanning participant enrolment and were validated in separate samples. The median positive predictive value was 57.0% (range 55.6% to 58.6%) and the median sensitivity was 4.3% (3.9% to 4.7%), calculated using a risk threshold of 0.5. The area under the receiver operating characteristic curve had a median value of 0.698 (range 0.688 to 0.701).

In relation to diagnoses of health conditions and other baseline variables, intervention patients differed markedly from the general population of the control areas (Table 1; Table 2). For example, intervention patients had 1.2 chronic health conditions on average, compared with 0.3 conditions for the general population. However, after matching, matched controls and intervention patients had similar characteristics. For example, both groups had 1.2 chronic health conditions on average, mean age of 65.5 years, mean predictive risk score of 0.17, similar prevalence of health conditions, and similar previous hospital use (Table 1; Table 2; Figure 3). Standardised differences were much lower than the 10% threshold, apart from diagnoses for angina (11.1%) and mental health conditions (15.0%).

Table 1: Balance, before and after matching, in the telephone health coaching study (demographic and health characteristics)

	Potential controls ³⁵ (n=969,677)	Intervention participants (n=2,698)	Matched controls (n=2,698)	Standardised difference (variance ratio)	
				Before matching	After matching
Age	41.6 (25.3) ³⁶	65.5 (13.4)	65.5 (13.5)	18.0 (0.28)	0.2 (0.98)
Female (%)	54.7 (n=522,049) ³⁷	47.7 (n=1,288)	47.7 (n=1,288)	14.0	0.0
Socioeconomic deprivation score ³⁸	34.4 (17.5)	37.2 (18.9)	36.8 (19.3)	14.8 (1.21)	1.8 (1.03)
Anaemia (%)	1.7 (n=16,797)	3.2 (n=86)	3.7 (n=99)	9.4	2.6
Angina (%)	2.3 (n=22,092)	15.4 (n=416)	11.6 (n=314)	47.6	11.1
Asthma (%)	3.8 (n=36,971)	5.7 (n=154)	6.9 (n=186)	8.9	4.9
Atrial fibrillation & flutter (%)	2.3 (n=22,370)	6.1 (n=165)	6.5 (n=176)	19.0	1.7
Cancer (%)	5.3 (n=51,530)	6.1 (n=165)	8.0 (n=216)	3.5	7.4
Cerebrovascular disease (%)	1.8 (n=17,807)	4.2 (n=112)	4.1 (n=111)	13.6	0.2
Congestive heart failure (%)	1.6 (n=15,055)	5.7 (n=154)	5.5 (n=149)	22.4	0.8
COPD (%)	1.9 (n=18,217)	5.6 (n=152)	5.5 (n=149)	19.8	0.5
Diabetes (%)	4.0 (n=38,805)	29.6 (n=799)	29.3 (n=791)	72.9	0.7
History of falls (%)	2.3 (n=22,549)	3.2 (n=85)	2.4 (n=66)	5.1	4.3
History of injury (%)	7.6 (n=73,450)	7.4 (n=201)	7.9 (n=213)	0.5	1.7
Hypertension (%)	8.7 (n=84,092)	28.4 (n=767)	31.2 (n=842)	52.6	6.1
Ischaemic heart disease (%)	4.0 (n=38,799)	19.9 (n=537)	19.7 (n=532)	50.6	0.5
Kidney failure (%)	1.1 (n=10,996)	2.9 (n=78)	2.9 (n=77)	12.5	0.2
Mental health conditions (%)	3.0 (n=28,910)	2.4 (n=64)	5.2 (n=141)	3.8	15.0
Peripheral vascular disease (%)	2.1 (n=20,735)	4.1 (n=111)	5.0 (n=134)	11.4	4.1
Number of long-term conditions	0.3 (0.8)	1.2 (1.5)	1.2 (1.5)	74.6 (3.41)	0.7 (1.02)
Predictive risk score	0.10 (0.08)	0.17 (0.12)	0.17 (0.2)	67.3 (2.15)	0.0 (1.00)

Note: Data are proportion (%) of individuals or mean (standard deviation) unless otherwise stated.

³⁵ Residents of control areas with previous hospital use.

³⁶ For complete cases (n=968,659).

³⁷ For complete cases (n=954,282).

³⁸ Taken from the Index of Multiple Deprivation 2007 (Communities and Local Government 2008).

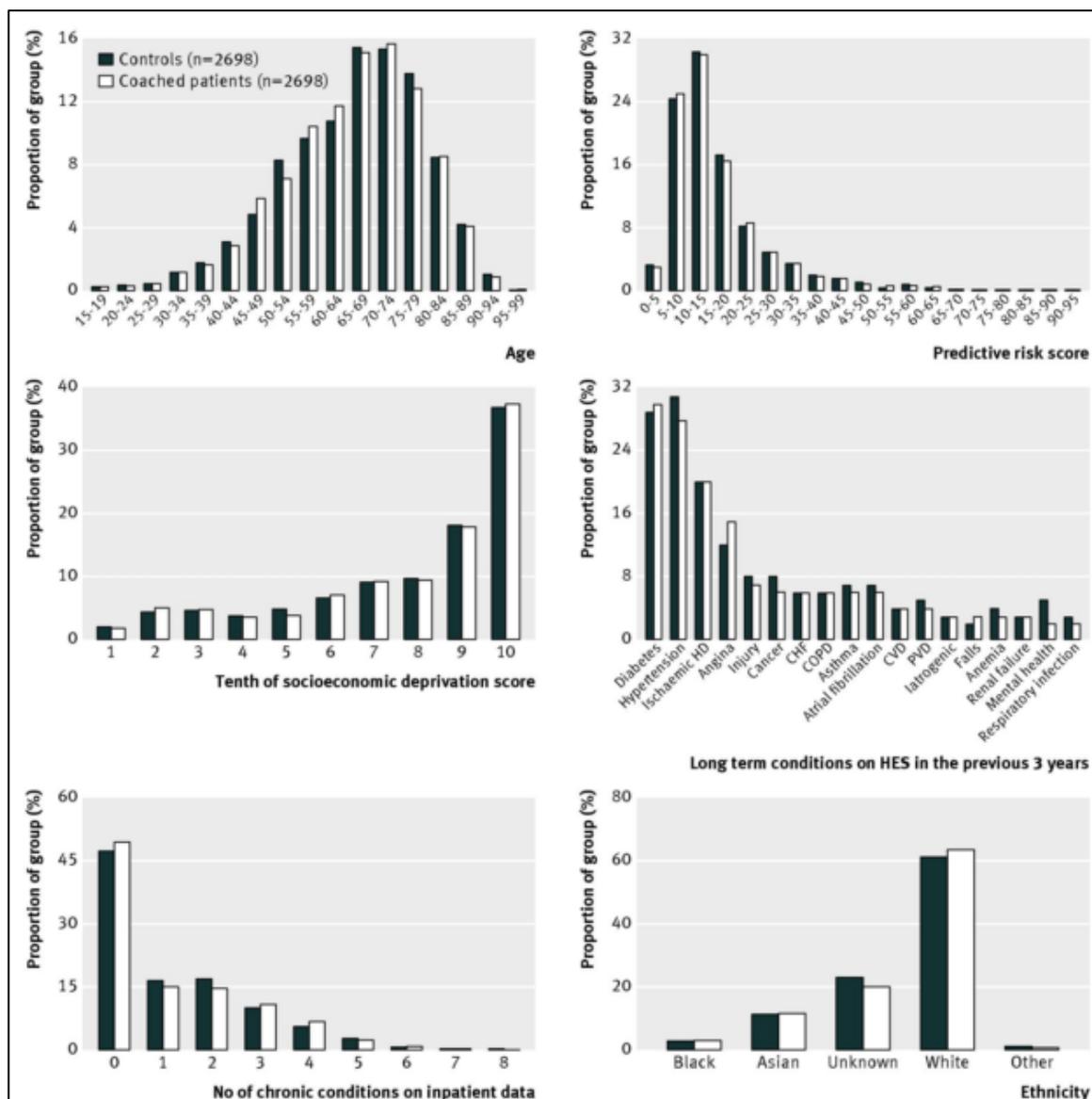
Table 2: Balance, before and after matching, in the telephone health coaching study (use of secondary care)

	Potential controls ³⁹ (n=969,677)	Intervention participants (n=2,698)	Matched controls (n=2,698)	Standardised difference (variance ratio)	
				Before matching	After matching
Emergency admissions (previous year)	0.2 (0.6)	0.3 (0.7)	0.3 (0.7)	19.7 (1.53)	2.1 (1.14)
Emergency admissions (previous month)	0.01 (0.13)	0.03 (0.21)	0.03 (0.17)	9.7 (2.37)	3.2 (1.49)
Elective admissions (previous year)	0.3 (1.4)	0.4 (1.1)	0.3 (0.9)	4.8 (0.56)	2.9 (1.44)
Elective admissions (previous month)	0.03 (0.21)	0.03 (0.19)	0.03 (0.19)	2.2 (0.81)	0.8 (1.02)
Outpatient attendances (previous year)	1.6 (3.0)	3.9 (5.0)	3.6 (4.5)	55.5 (2.66)	6.1 (1.24)
Outpatient attendances (previous month)	0.15 (0.54)	0.36 (0.75)	0.33 (0.72)	32.8 (1.88)	3.9 (1.07)
Emergency bed days (previous year)	1.06 (7.64)	1.80 (8.02)	1.57 (6.16)	9.5 (1.10)	3.3 (1.70)
Emergency bed days (previous year, trimmed to 30 days)	0.77 (3.74)	1.44 (4.79)	1.38 (4.58)	15.6 (1.64)	1.3 (1.09)

Note: Data are mean (standard deviation) unless stated otherwise.

³⁹ Residents of control areas with previous hospital use.

Figure 3: Balance after matching in the telephone health coaching study (selected variables, n=2,698 coached patients and 2,698 matched controls)



Notes: HD=heart disease; CHF=congestive heart failure; COPD=chronic obstructive pulmonary disease; CVD=cerebrovascular disease; PVD=peripheral vascular disease. *Score of 10=most deprived; tenths were defined on the basis of national data for the Index of Multiple Deprivation 2007 (Communities and Local Government 2008).

3.5.2 Comparing hospital use and costs

Intervention patients had more emergency admissions in the year after enrolment than in the year before enrolment (0.38 *vs.* 0.31 admissions per head). A smaller increase was observed for matched controls (Table 3, Figure 4). Comparing the two groups, emergency admissions increased by 0.05 per head more among intervention patients than among matched controls (95% confidence interval 0.00 to 0.09, $p=0.046$; Table 4), which was a relative increase of 13.6% (0.2% to 27.1%).

Table 3: Rates of secondary care use before and after enrolment into telephone health coaching, and equivalent figures for matched controls

	Participants (n=2,698)			Matched controls (n=2,698)		
	Year before enrolment	Year after enrolment	Difference	Year before enrolment	Year after enrolment	Difference
Emergency admissions	0.31 (0.74)	0.38 (0.93)	0.07* (0.96)	0.29 (0.69)	0.32 (0.88)	0.03 (0.95)
Bed days (after emergency admission)	1.80 (8.02)	2.90 (11.85)	1.09* (12.86)	1.57 (6.16)	2.57 (10.79)	1.00* (11.29)
Elective admissions	0.37 (1.05)	0.39 (1.10)	0.02 (1.20)	0.34 (0.87)	0.38 (1.11)	0.04 (1.25)
Outpatient attendances	3.90 (4.96)	3.96 (5.13)	0.06 (4.29)	3.61 (4.45)	3.30 (4.91)	-0.30* (4.28)
Secondary care costs (£)	1388 (2683)	1650 (3228)	261* (3534)	1292 (2436)	1379 (3054)	86 (3355)

Notes: Data are mean number or cost per head (standard deviation). *Difference= $p<0.05$

Figure 4: Rates of emergency hospital admissions before and after enrolment into telephone health coaching, and equivalent figures for matched controls

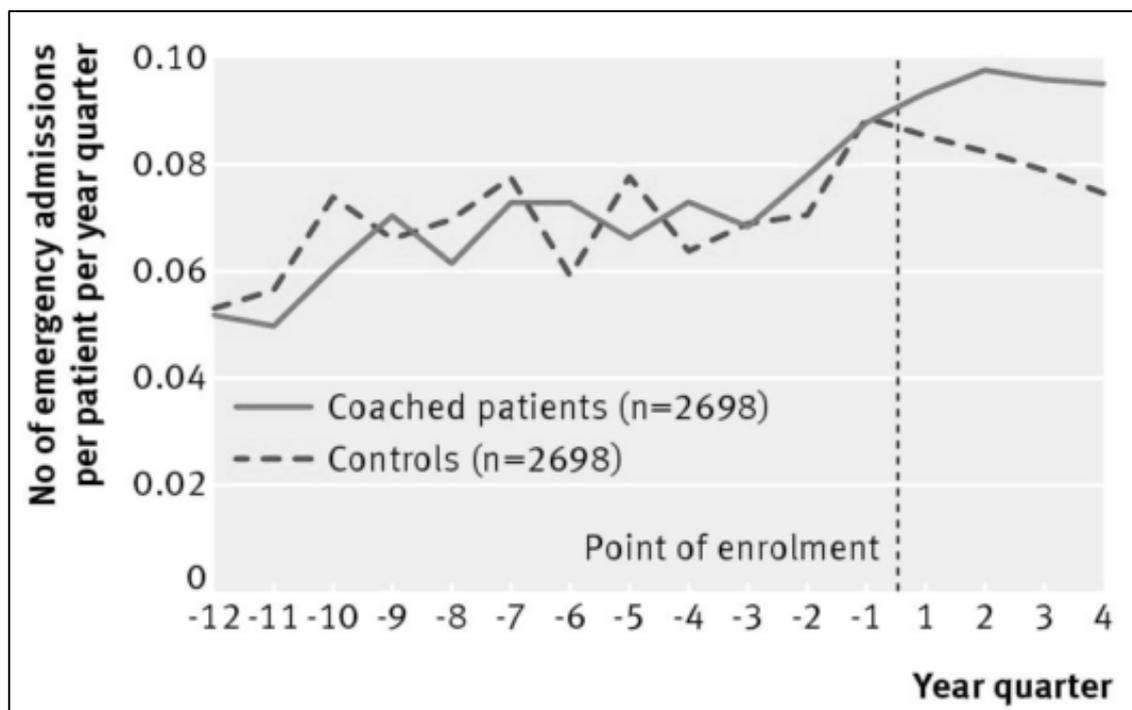


Table 4: Estimated treatment effects of telephone health coaching on secondary care use

	Mean (standard deviation)	95% confidence interval	p-value
Emergency admissions	0.05 (1.20)	0.00 to 0.09	0.046
Bed days (after emergency admission)	0.09 (15.87)	-0.50 to 0.69	0.757
Elective admissions	-0.02 (1.48)	-0.07 to 0.04	0.549
Outpatient attendances	0.37 (5.50)	0.16 to 0.58	<0.001
Secondary care costs (£)	175 (4592)	22 to 328	0.025

Note: Table shows rates per head.

Differences between groups in hospital bed days and elective admissions were not significant (Table 4). Outpatient attendances rose by 0.37 attendances per head more among intervention patients than among matched controls (95% confidence interval 0.16 to 0.58, $p < 0.001$), due to a fall among controls. Overall secondary care costs increased by £175 (€203; \$268) per head more among intervention patients than among controls (£22 to £328, $p = 0.025$).

3.5.3 Sensitivity analysis for unobserved confounding

Within 12 months of the intervention, 2.3% of patients in both the intervention (n=63) and matched control (n=63) groups died in hospital. Sensitivity analysis simulated a hypothetical unobserved confounding variable and showed that, for the apparent increase in emergency admissions to be reversed, such a variable would need to be strongly associated with both intervention status and outcome, with odds ratios greater than 2.8. By comparison, insulin treatment (which is one variable we did not observe) had an odds ratio of 1.6 with intervention status (Jordan, Lancashire, and Adab 2011).

3.6 Discussion

3.6.1 Statement of findings

Telephone health coaching aims to support patients in managing their long-term health conditions. The hope is that, by promoting healthy behaviours and by providing a means to identify problems before they become critical, telephone health coaching can help prevent crises that lead to hospital admissions. We compared a large sample of people receiving telephone health coaching in England with a well-balanced, retrospectively matched control group using person-level data. Rather than see a reduction in hospital activity in the study group, we found that emergency admissions increased at a faster rate among intervention patients than matched controls, as did outpatient attendances and secondary care costs. Therefore, there was no evidence of reductions in hospital admissions, and no savings were detected from which to offset the cost of the intervention.

3.6.2 Strengths and weaknesses

We were able to study a large number of intervention patients with a high rate of data linkage (87%). Imperfect linkage was mainly due to imperfect recording of individual identifiers on the service's operational system, because most records that did not link had missing or incomplete personal linkage data. On the assumption that recording omissions happened at random, our sample was an unbiased sample from the population receiving the intervention. Although the analysis then focussed on patients with previous hospital use, this cohort is where the scope for savings was highest.

The use of administrative data meant that data were available for a high proportion of patients, and avoided problems of under-reporting by patients about how many services were used (Richards, Coast, and Peters 2003). However, the quality of data was not directly under our control. Potential problems with administrative data included limited insight into the quality and appropriateness of care (Roos et al. 1993), and observational intensity bias if coding practices varied between geographic areas (Wennberg et al. 2013).

We obtained data for more patients than our sample size calculation suggested was needed (2,698 *vs.* 2,035). Therefore, although we originally envisaged that we would only be able to detect differences in emergency admission rates of 15% or higher, the 13.6% increase detected was statistically

significant, and was unlikely to be the result of chance ($p=0.046$). A 13.6% increase in emergency admissions is substantial for the health service, and much more than the general increase in age standardised rates of admission of 2.5% a year (Bardsley et al. 2013).

The main risk to validity in this observational study was that, although intervention and matched control groups were similar in terms of an established set of predictors of future hospital use, they could have differed in ways that we could not observe (that is, there may have been unobserved confounding). Typically, only a small proportion of eligible patients receive complex interventions out of the hospital (Subramanian et al. 2004). Birmingham OwnHealth was a relatively established service, and at least 80% of the local general practices had participating patients. Nevertheless, there are around 9,000 patients in the area with uncontrolled diabetes (The NHS Information Centre for health and social care 2012), for example, while only around 3,000 patients received the intervention in the time period chosen.

We sought to minimise unobserved confounding by careful selection of the pool of potential controls, matching on previous outcomes and difference-in-difference estimation. The eligibility criteria for the service included clinical variables such as HbA1c. These variables were not recorded in our dataset. However, we ensured that the prevalence of health conditions was similar between the two groups, as were variables that are correlated with clinical indicators, such as hospital use (Govan et al. 2011). Sensitivity analysis showed that, although the increase in emergency admissions could conceivably have been caused by unobserved confounding, it is unlikely that we missed a reduction. To have missed such a reduction, the amount of unobserved confounding would have had to be greater than is realistic for clinical variables. Further, it is reassuring that no differences were observed in in-hospital mortality between the two groups. For example, if disease control had been worse among intervention patients, more deaths might have been expected.

Observational study designs have some advantages over RCTs. This study looked at a population that was selected to participate in telephone health coaching in routine practice. By contrast, RCTs may have poor generalisability when the patients in the trial differ from those who would receive the intervention routinely. Such differences might occur because of the selection of health care settings or practitioners for a trial, the choice of eligibility criteria for a trial, certain individuals preferring not to participate, or study design (Rothwell 2005).

This study investigated effects on hospital use and associated costs for people enrolled between 2006 and 2008. The effect of the service might have changed over time, because the eligibility criteria were later broadened to include chronic kidney disease, stroke, and transient ischaemic attack from 2009; and hypertension and older patients at risk from 2010. Although not the focus of this study, Birmingham OwnHealth might have affected the use of primary care or health-related quality of life. Because the median duration of enrolment was several years, effects could have been over longer

time periods than those analysed in this study. Other work has found improvements in clinical metrics among participants with poorly controlled diabetes (Jordan, Lancashire, and Adab 2011), as well as high levels of patient satisfaction (Azarmina and Lewis 2007).

3.6.3 Comparison with other studies

Previous studies of the effect of telephone health coaching on service use have generally been encouraging. Four of nine studies identified by a systematic review found evidence of an effect on health service use, although sample sizes were typically small (Hutchison and Breckon 2011). A more recent, large RCT found reductions in hospital admissions and expenditures (Wennberg et al. 2010). Before the current study, the largest observational study of telephone health coaching included 874 Medicaid members, and found no effect (Lin et al. 2012). The current study supports their findings on a larger sample drawn from England (n=2,698).

One possible explanation for the apparently contradictory nature of the findings from these studies might be subtle differences in the design of the interventions. Aspects of intervention design such as the frequency of telephone calls vary widely between studies (Hutchison and Breckon 2011), although the profile of telephone calls in the current study (usually monthly) was not out of line. The largest RCT (Wennberg et al. 2010) included decision making for preference sensitive conditions, while three of the four effective interventions identified by the systematic review involved telemonitoring of vital signs in addition to health coaching. Telemonitoring could be effective at reducing hospital admissions even when combined with automated motivational messages and symptom questions rather than health coaching (Steventon et al. 2012a).⁴⁰ This study adds weight to the conclusions suggested by the Medicaid studies (Lin et al. 2012; Kim et al. 2013) that health coaching is not effective at reducing hospital use by itself. Further, although potentially explained by unobserved confounding, we found evidence that the intervention in Birmingham increased emergency admissions. Previous evaluations of other complex interventions out of hospital have also found indications of increases (Gravelle et al. 2007). One possibility is that increases occur as a result of greater observation. Indeed, in other settings, more intense observation and greater use of diagnostic tests have been found to correlate with the number of medical interventions made (Lucas et al. 2008).

Discrepancies between the findings of different studies could also be due to study settings, with both the Medicaid studies and the current study relating to a publicly insured population. Targeting the intervention using the outputs from a predictive risk model may increase effectiveness, as could better integration with existing primary and secondary care services. Finally, the evaluation method could affect results. Confounding is possible in observational studies, although our study design attempted to limit this threat to validity as much as possible.

⁴⁰ This is the study in Chapter 5.

3.6.4 Conclusions

We conclude that Birmingham OwnHealth did not lead to the anticipated levels of reductions in hospital admissions or associated costs. Based on a systematic review and subsequent studies, including the present study, standard telephone health coaching seems unlikely to lead to reductions in hospital use, without the addition of other elements such as telemonitoring, shared decision making for preference sensitive conditions, or predictive modelling. More care coordination might also be needed. Unless health coaches have established relationships with other clinical staff, new interventions could prove to be additions to existing patterns of service use, rather than create efficiencies.

The study serves as a warning that efficacy as demonstrated by RCTs might not imply effectiveness in routine practice (Haynes 1999). Because administrative datasets are regularly updated, the methods used in the present study may be useful to monitor new services to ensure that benefits are achieved.

Box 3: What this study adds

What is already known on this topic

Telephone health coaching provides support and encouragement to patients to manage long-term health conditions. It is hoped that hospital admissions will be prevented as a result, creating efficiency gains for health care systems. However, the current evidence base is unclear; many studies have been small and interventions are heterogeneous.

What this study adds

This study adds weight to the existing view that health coaching by itself is not effective at reducing hospital use over 12 months. Coaching could be coupled with other interventions such as shared decision making or telemonitoring, and the context in which interventions are delivered might also be crucial. Efficacy of new services as demonstrated by RCTs might not imply effectiveness in routine practice.

Chapter 4 Comparison of strategies to choosing the control population (research paper 2)

4.1 Preamble to research paper 2

In the case study presented in Chapter 3, matched controls were selected from areas of England that were similar to the intervention area in terms of demography, occupational mix, and rates of education, occupation, limiting long-term illness and unpaid care. This was done partly because the large size of the OwnHealth service was thought to reduce the scope to find matched controls locally. However, other studies have taken different approaches to control area selection, including local (McConnell et al. 2008) and national controls (Roland et al. 2012).

The choice of the population from which to select controls will in general influence confounding at both the individual and area levels. The next paper compares the effects of alternative strategies for the choice of control area on the balance achieved by person-level matching algorithms, and reports the relative bias and mean-squared error of the resulting estimates of treatment effect. It begins with a reanalysis of a Partnership for Older People Projects (POPPs) study, which concerned the effect of a rapid response service on hospital admissions. Like the Birmingham OwnHealth study, the POPPs study chose controls from within matched areas of England (Steventon et al. 2011, 2012b). The next paper additionally obtains controls from within the intervention area and nationally. The case study was used to calibrate a series of simulations that could assess bias and mean-squared error directly.

The paper elucidates some situations in which lower bias will be obtained when selecting controls from within matched areas than when selecting controls locally. However, these situations are fairly restrictive. An important consideration is that there is a relatively large amount of unexplained variation in hospital admissions between areas of England. This will mean that, when addressing this endpoint, local controls will generally lead to lower bias than the other approaches.

This paper has been published in *Health Services Research and Outcomes Methodology* (Steventon, Grieve, and Sekhon 2015) under the terms of the Creative Commons Attribution License, which permits any use, distribution, and reproduction in any medium, provided the original author(s) and the source are credited. Minor formatting and stylistic changes have been made to the version that was accepted for publication.

RESEARCH PAPER COVER SHEET

PLEASE NOTE THAT A COVER SHEET MUST BE COMPLETED FOR EACH RESEARCH PAPER INCLUDED IN A THESIS.

SECTION A – Student Details

Student	Adam Steventon
Principal Supervisor	Professor Richard Grieve
Thesis Title	Evaluating complex interventions using routinely collected data: Methods to improve the validity of randomised controlled trials and observational studies

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?	Health Services and Outcomes Research Methodology		
When was the work published?	2015		
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?*	No	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	Choose an item.

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)	I contributed to the specification of the research question and proposed the study design for the case study analysis. I conducted all analyses, including the simulations, and produced the first draft of the manuscript, before incorporating comments from the co-authors and peer reviewers.
--	---

Student Signature: _____

Date: _____

Supervisor Signature: _____

Date: _____

4.2 Abstract

Various approaches have been used to select control groups in observational studies: 1) from within the intervention area; 2) from a convenience sample, or randomly chosen areas; 3) from areas matched on area-level characteristics; and 4) nationally. The consequences of the decision are rarely assessed but, as we show, it can have complex impacts on confounding at both the area and individual levels.

We began by reanalysing data collected for an evaluation of a rapid response service on rates of emergency hospital admission. Balance on observed individual-level variables was better with external than local controls, after matching. Further, when important prognostic variables were omitted from the matching algorithm, imbalances on those variables were also minimised using external controls. Treatment effects varied markedly depending on the choice of control area, but in the case study the variation was minimal after adjusting for the characteristics of areas.

We used simulations to assess relative bias and means-squared error, as this could not be done in the case study. A particular feature of the simulations was unexplained variation in the outcome between areas. We found that the likely impact of unexplained variation for hospital admissions dwarfed the benefits of better balance on individual-level variables, leading us to prefer local controls in this instance. In other scenarios, in which there was less unexplained variation in the outcome between areas, bias and mean-squared error were optimised using external controls. We identify some general considerations relevant to the choice of control population in observational studies.

4.3 Introduction

Well-conducted randomised controlled trials (RCTs) are often considered the gold standard in comparative effectiveness research as they can balance both observed and unobserved variables between treatment groups. However, for many examples in health services and outcomes research, RCTs are infeasible and the best available information on effectiveness comes from an observational study. These must be designed and analysed carefully so that findings are not biased by differences in the characteristics of patients or settings (Rubin 2010). While techniques such as instrumental variable estimation can handle confounding due to unobserved as well as observed characteristics (Stukel et al. 2007), valid instruments are rare, so instead studies tend to use approaches that assume no unobserved confounders. For example, propensity scores can be used to select, from a wider population of potential controls, a matched subgroup that is similar to the intervention group with respect to observed variables (Rosenbaum and Rubin 1983). Matching methods are appealing because, for some estimands, regression models are more robust to model specification when applied to matched rather than unmatched data (Ho et al. 2007).

Many advances have been made in analytical methods for observational studies. For example, genetic matching uses computer-intensive search algorithms to find more closely balanced matched control groups than traditional approaches using the propensity score (Sekhon and Grieve 2012). Also, doubly robust methods can provide unbiased estimates when either the treatment selection or the outcome model is correctly specified (Bang and Robins 2005). On the other hand, relatively little attention has been paid to study design (Rubin 2007). This is an important omission as improvements in design could reduce the main threat to the validity of observational studies, namely confounding due to unobserved variables. One design issue that has received scant attention relates to the choice of higher-level unit from which the control group is selected. The issue arises because interventions are often piloted within a sample of units, such as hospitals, geographic areas, or schools. Within these units, a subset of individuals will receive the intervention and others the control (*e.g.*, usual practice). However, often data are available for other potential control units. Therefore, investigators have a choice at the design stage of a study between selecting matched controls from within the intervention areas (or, more generally, higher-level units), from other areas, or nationally. If selecting controls from other areas, these units could be selected as part of a convenience sample, or matched to the characteristics of the intervention areas. Studies have used the full range of approaches (McConnell et al. 2008; Nelson 2012; Roland et al. 2012).

In theory, the choice of control area can have complex implications for confounding at both the individual and area levels. While selecting controls from within an intervention area will automatically give perfect balance on area-level variables, it will not always give good balance on individual-level variables, as there may be limited overlap between the characteristics of treated and untreated individuals in the intervention area (Stuart and Rubin 2008). In situations of limited overlap, selecting

controls externally may give better balance on both observed and unobserved individual-level variables than selecting controls from within the intervention area. To illustrate this hypothesis, suppose that, within the intervention area, a particular characteristic (older age) is associated with treatment receipt. In this situation, the local untreated group will contain relatively few older people, as these have been disproportionately recruited into the intervention. By contrast, there may be a relatively high number of older people in areas not offering the intervention, making it easier to obtain good matches on that variable when using external rather than local controls. The same argument applies to unobserved variables (such as extent of social support), but because unobserved variables cannot be taken into account by commonly used analytical approaches, they are particularly important to balance by design. Based on these considerations, we might expect the optimal strategy for selecting control areas to depend on the extent of confounding at the individual level versus area level. This reasoning was used by Griswold and Localio (2010) in an observational analysis of the effect of concurrent use of proton-pump inhibitors and clopidogrel, and led them compare an approach using local controls with approaches using national controls and controls from similar hospitals. They found that local controls produced worse balance on observed individual-level variables, but they could not assess unobserved variables, bias, or statistical efficiency.

While careful selection of control areas has long been recognised as crucial for case-control studies (Miettinen 1985), little methodological research has been undertaken to guide the choice of control population in cohort studies, which are the focus of this paper. Meta-epidemiological work has found that observational studies tend to give treatment effects that are more similar to those from RCTs when their control group is sourced locally rather than from a matched area (Glazerman, Levy, and Myers 2003). Discrepancies in treatment effects tend to be larger still when the observational study uses a convenience sample of areas or takes controls from a national sample. Meta-epidemiological work can control for only a limited number of study characteristics (Deeks et al. 2003), but some studies have controlled for research setting more closely by comparing randomised and observational studies that share a single treatment group. A review of these also concluded that local controls should be preferred (Cook, Shadish, and Wong 2008), though the number of constituent studies was small and dominated by labour market interventions (Shadish, Clark, and Steiner 2008). These reviews did not examine whether the relative benefit of local versus external control groups varies between alternative research settings with different levels of confounding at the local and area levels. Furthermore, although RCTs are often considered to be the gold standard in comparative effectiveness research, in some cases there are legitimate reasons why observational studies should give different treatment effects to RCTs, relating to measurement, study samples, or interventions (Hartman et al. 2015).

Several research methods use multiple control groups to assess bias, but these do not address the prior question about which control area should be preferred. For example, Campbell (1969) proposed

using multiple control groups to put bounds on treatment effects, and also to confirm that the variation in treatment effects is as expected given prior information about the different groups (this approach is known as 'control by systematic variation'). Rosenbaum (1987) gave a detailed account of how multiple control groups could be used to test for unobserved confounding at the individual level, though he did not consider area-level confounding. Lu and Rosenbaum (2004) considered the situation of one treatment group and two control groups, and developed a matching algorithm to make three pairwise comparisons while optimising the use of individuals. Multiple control groups are sometimes used in the medical literature, for example by comparing treated patients to both historic and concurrent controls (Harrison et al. 2010), but they seem to be rare in comparison to studies that use a single group (Austin 2008a).

As few studies have assessed the implications of the choice of control population on bias and Mean-Squared Error (MSE), we conducted a simulation study. We calibrated these simulations to a case study of an intervention that aimed to reduce emergency hospital admissions for older people, and then varied the assumptions made for the propensity score model and the response model across a total of 45 scenarios. Although we expected local controls to be preferable in many scenarios, the simulation reports that in some settings a strategy of selecting external controls gives lower bias and lower MSE. This paper provides general methodological recommendations to inform the choice of control populations in future studies. We also append R code to provide practical tools for investigators to undertake simulations at an early stage of study design to determine which control population is likely to give the least bias in their research setting. These simulations could complement the use of multiple control groups, if these are available, by identifying which control group is likely to give the most reliable inferences.

This paper is organised as follows. First, we describe the estimands typically of interest in a cohort study and the various ways in which the choice of control area can affect bias. Then, in Section 4.5, we describe the case study and show how balance and estimated treatment effects depend on the choice of control area. Section 4.6 describes the simulation design and sets out a number of scenarios with varying levels of individual-level and area-level confounding. The results of the simulation study are given in Section 4.7, and the final section concludes.

4.4 Statistical considerations relating to the choice of control area

Following the Rubin causal model (Rubin 1978), we begin by positing two potential outcomes for each individual, $Y(1)$ and $Y(0)$, relating to outcomes under intervention and control, respectively. A common target estimand is the Sample Average Treatment effect for the Treated (SATT), defined as the average difference between these potential outcomes over the group of people receiving the intervention (Imai, King, and Stuart 2008). Common approaches to estimate the SATT assume no

unobserved confounding. In other words, if Z indicates assignment to the treatment, then we assume that there is a set of observed baseline variables X such that:

$$[1] \quad Y(1), Y(0) \perp\!\!\!\perp Z \mid X$$

If this assumption is valid, then an unbiased estimate of the treatment effect at each level of X can be obtained by calculating the differences in the observed responses of intervention and control patients at that level (Rosenbaum and Rubin 1983). Therefore, an unbiased estimate of SATT can be produced by selecting a matched control group with the same distribution of X as the intervention group. This can be accomplished by matching on a single scalar quantity known as the propensity score, which reflects the probability of treatment assignment, conditional on observed variables (Rosenbaum and Rubin 1983). A complementary approach, genetic matching, uses a computer-intensive search algorithm to find the matches that maximise balance across all variables in X , given the data (Sekhon and Grieve 2012).

Bias can arise when matching for several reasons. First, some important baseline variables may be unobserved, and so omitted from the matching algorithm. Second, the matching algorithm may not produce groups with the same distribution of observed variables. We decompose the estimation error into these two components by supposing an additive model for the outcome of each individual (Imai, King, and Stuart 2008):

$$Y_i(t) = g_t(X_i) + h_t(U_i)$$

Here, g_t and h_t are, in general, unknown functions ($t = 0, 1$), and U represents the unobserved baseline variables. Following 1-1 matching, the estimation error is:

$$SATT - D = \frac{1}{n} \sum_{i \in \{i|Z_i=1\}} \{g_1(X_i) + h_1(U_i) - g_0(X_i) - h_0(U_i)\} \\ - \left\{ \frac{1}{n} \sum_{i \in \{i|Z_i=1\}} \{g_1(X_i) + h_1(U_i)\} - \frac{1}{n} \sum_{i \in \{i|Z_i=0\}} \{g_0(X_i) + h_0(U_i)\} \right\}$$

or:

$$\frac{1}{n} \sum_{i \in \{i|Z_i=0\}} \{g_0(X_i) + h_0(U_i)\} - \frac{1}{n} \sum_{i \in \{i|Z_i=1\}} \{g_0(X_i) + h_0(U_i)\}$$

Here, D is the estimator, and n is the number of individuals in the intervention group, which is the same as the number of individuals in the matched control group because the matching is 1-1.

The set of observed baseline variables X contains some variables at the individual level ($X_{i,1}$) and some at the area level ($X_{i,2}$), and likewise for the unobserved baseline variables, U . We assume that the outcome model can be further decomposed into additive subcomponents relating to observed and unobserved variables at the individual and area levels, and thus write, for example:

$$g_t(X_i) = g_t^1(X_{i,1}) + g_t^2(X_{i,2}).$$

Where g_t^1 and g_t^2 are, in general, unknown functions that represent the subcomponents of the outcome model relating to observed variables at the individual and area levels, respectively ($t = 0, 1$), and likewise for the unobserved variables at these levels, h_t^1 and h_t^2 .

With this decomposition, there are four terms to the estimation error, representing the effects of imbalance on observed individual-level variables, unobserved individual-level variables, observed area-level variables, and unobserved area-level variables, respectively:

$$\frac{1}{n} \sum_{i \in \{i|Z_i=0\}} g_0^1(X_{i,1}) - \frac{1}{n} \sum_{i \in \{i|Z_i=1\}} \{g_0^1(X_{i,1})\}$$

$$\frac{1}{n} \sum_{i \in \{i|Z_i=0\}} h_0^1(U_{i,1}) - \frac{1}{n} \sum_{i \in \{i|Z_i=1\}} \{h_0^1(U_{i,1})\}$$

$$\frac{1}{n} \sum_{i \in \{i|Z_i=0\}} g_0^2(X_{i,2}) - \frac{1}{n} \sum_{i \in \{i|Z_i=1\}} \{g_0^2(X_{i,2})\}$$

$$\frac{1}{n} \sum_{i \in \{i|Z_i=0\}} h_0^2(U_{i,2}) - \frac{1}{n} \sum_{i \in \{i|Z_i=1\}} \{h_0^2(U_{i,2})\}$$

We now consider the situation in which there is one intervention area (containing both treated and untreated individuals) and several external areas (containing only untreated patients). Many of the issues would be the same if there were multiple intervention areas, a point we return to in Section 4.8.

If controls are selected from within the intervention area (forthwith referred to as 'local controls'), then the third and fourth error terms will be zero, as both observed and unobserved area-level variables will be balanced automatically by design. However, the first error term will be nonzero when the matching algorithm is not able to balance the observed variables, $X_{i,1}$. This can happen when there is poor overlap between the characteristics of treated and untreated individuals, as might be expected when the $X_{i,1}$ are strong confounders (*i.e.*, when they are strongly predictive of treatment assignment and outcome) or when a high proportion of local individuals receive the treatment (*i.e.*, high intervention saturation). The second error term will in general be nonzero because matching

algorithms cannot balance unobserved variables, except to the extent to which they are correlated with the observed variables that the matching algorithm is required to balance.

We explore three other strategies that may produce better balance on the individual baseline variables than local controls. These three other strategies use controls external to the intervention area, so the third and fourth error terms may be nonzero. However, individuals with the characteristics likely to lead to treatment assignment in the intervention area will not have been lost from the supply of potential controls in external areas. Thus, all other things being equal, overlap will be higher when controls are selected externally rather than locally. In turn, this can result in smaller errors through terms one and two.

Strategy 2 represents a commonly used, but perhaps ill-advised, approach whereby control areas are selected as part of a convenience sample, with little attention placed on the characteristics of the control areas. In the case study that follows, we implement this strategy a large number of times to show the substantial variability that arises in terms of balance and treatment effects. Strategy 3 is a better approach whereby the control area is matched to the intervention area with respect to the observed area-level variables $X_{i,2}$, thus minimising the third error term, but not necessarily the fourth error term. The final strategy (strategy 4) is a national approach in which controls are taken from all areas external to the intervention area (*i.e.*, a national sample). This approach maximises the number of potential controls, which may result in closer matches on individual-level variables at the expense of worse balance on area-level variables than strategy 1.

The subsequent case study and simulation contrast these strategies across situations that typically arise in health services and outcomes research.

4.5 Case study: rapid response service for older people

A rapid response service was introduced into a large, rural, district county area of England, as part of a national program to improve partnership working across care sectors (the Partnership for Older People Projects, or POPPs) (Windle et al. 2009). An important objective of the rapid response service was to prevent emergency hospital admissions for older people through prompt treatment close to the individual's home. However, a previous evaluation using controls from matched geographic areas found that the service had the opposite effect and increased these admissions, perhaps because of additional health needs identified by the rapid response team (Steventon et al. 2011, 2012b).

We focus on the subset of older people (aged 70 or over) who were enrolled into the rapid response service during October 2008 with a history of hospital admissions ($n=108$). We examine the effect of the service on the likelihood of participants having one or more emergency hospital admission during the 12 months following enrolment. As in the original study, we obtained individual-level variables from the Hospital Episode Statistics (HES), which is a national database containing details of all

hospital care funded by the National Health Service in England. Unlike the original study, we obtained these variables for people in each of the 33 district counties in England. Thus, we were able to apply the various strategies for control area selection described in Section 4.4.

Individual-level variables included: age; sex; socioeconomic deprivation score (defined at a small-area level⁴¹); diagnoses of four specific health conditions; total number of chronic health conditions; numbers of prior elective and emergency hospital admissions; and a predictive risk score. The predictive risk score was an estimate of the probability of one or more emergency hospital admission during the 12 months following enrolment, under usual care. It was based on an existing predictive risk model, with coefficients reweighted to match the patterns of hospital use that we observed for untreated individuals in the intervention area (Billings et al. 2006). We applied these reweighted coefficients to people in the other district counties to calculate their risk scores. Of all the variables, the predictive risk score, age and number of prior emergency hospital admissions were the most strongly predictive of the outcome.

Strategy 2 was repeated 32 times (once for each of the district counties in England, excluding the intervention area). The matched geographic area for strategy 3 was selected according to an established method that is used to produce comparative statistics (Office for National Statistics 2010). This involved minimising the Euclidean distance from the intervention area with respect to a standard set of 43 area-level variables, relating to: population age structure; population density; ethnic mix; average household size and structure; education; overall rates of long-term illness; transport; overall employment rates; and the prevalence of various occupations. The method was similar to that used to select the matched geographic area in the original study (Steventon et al. 2011), except that we used a wider set of variables. We implemented strategy 4 by pooling potential controls from all 32 counties.

For each strategy, we applied the study inclusion criteria to define a pool of potential controls who were aged 70 or over in October 2008 and had a history of hospital admissions. To remove the possibility that differing population sizes influenced results, we reduced the eligible population of each area to a random sample of 500. As 108 individuals received the intervention, this left 392 potential controls for strategy 1, producing a saturation of just less than 30%. The number of potential controls for strategies 2 and 3 was 500, while in strategy 4 it was 16,000.

Within each of the chosen areas, we selected matched controls at the individual level using genetic matching (Sekhon and Grieve 2012). Matched controls were selected on a 1-1 basis with replacement, as this will typically lead to better balance on individual-level variables than matching without

⁴¹ Socioeconomic deprivation score was defined at the 'Lower Super Output Area' level, consisting of around 1,600 people on average, across all ages. In comparison, the county areas contained around 700,000 people on average. Therefore, socioeconomic deprivation score was treated as an individual-level variable, and attributed to individuals based on their place of residence.

replacement, or 1-n matching ($n > 1$). Balance before and after matching was assessed using the standardised difference, defined as the difference in sample means as a proportion of the pooled standard deviation (Austin 2009a). Although covariate balance should ideally be maximised without limit, a standardised difference of more than $\pm 10\%$ has been used to denote meaningful imbalance (Normand et al. 2001). We report estimated treatment effects using the absolute risk difference (*i.e.*, difference in proportions) and the relative risk difference, together with 95% confidence intervals produced using methods that recognised the dependencies within matched data (Agresti and Min 2004).⁴²

4.5.1 Case study results

The three most prognostic variables (*i.e.*, predictive risk score, age and number of prior emergency admissions) had lowest standardised differences before matching when the control population was defined using a matched geographic area (strategy 3). For example, age had a standardised difference of 49.2% before matching in strategy 3, compared with 56.9% in strategy 1 (Table 5). By contrast, the socioeconomic deprivation score had lowest standardised difference when controls came from within the intervention area.

After matching, strategy 4 (national controls) gave the best balance across all individual-level variables, reflecting the larger population size. For example, age had a standardised difference of 2.7% under national controls (Table 5), compared with 4.9% when using local controls. However, the intervention area was different from the national sample in terms of area-level variables, such as the proportion of residents who were aged 65 or over (Figure 5). Although strategy 3 (matched area) reported higher standardised differences than strategy 4 at the individual level (*e.g.*, 4.8% *vs.* 2.7% for age), it nonetheless outperformed local controls (average standardised difference 4.0% *vs.* 4.9%). Furthermore, the matched area was more like the intervention area than the national sample (Figure 5).

As expected, estimated treatment effects from strategy 2 (convenience sample) were very sensitive to the area chosen, and relative risk ratios ranged from 1.35 (95% CI, 1.00 to 1.82) to 3.87 (2.43 to 6.16). However, estimated treatment effects for strategies 1, 3, and 4 were very similar, with relative risk ratios of 2.07, 1.93 and 2.07, respectively (Table 6).

4.5.2 Inducing unobserved confounding

As we hypothesised that the relative strengths of the strategies will depend on the extent of unobserved confounding at the individual level, we repeated the analysis after omitting two important prognostic variables from the genetic matching algorithm, namely age and predictive risk score.⁴³

⁴² Because we matched with replacement, an individual may have been selected as the control for several intervention patients. The 95% confidence intervals that we report do not allow for any dependency within the data from matching with replacement.

⁴³ As age is one of the constituent variables of the predictive risk score, it was necessary to omit both variables.

Thus, we treated these variables as being unobserved. After matching, standardised differences on these variables were high across all strategies, but they were lower when using a matched control area than when using local controls, reflecting the generally better balance that existed before matching. Standardised differences for age were 44.3% and 49.4%, after matching, in these two strategies, respectively.⁴⁴ Standardised differences for the predictive risk score were 27.4% and 31.3%, respectively.

As would be expected, each strategy reported higher estimated treatment effects when the two prognostic variables were omitted from the matching algorithm. However, these increased by less when controls were sourced from a matched control area than with local controls (Table 6).

The case study findings suggest that, all other factors being equal, using external control groups can lead to lower standardised differences on observed variables than local controls. Furthermore, estimated treatment effects were more robust to the unobserved confounding considered when matched controls came from a matched area rather than locally. Although external controls lead to better balance on individual-level variables, they also introduce the possibility of confounding due to area-level differences. We could not quantify the impact of this phenomenon in the case study, though the variability of the treatment effects observed under strategy 2 suggests it may be substantial. We now use simulations, calibrated to these data, to assess the implications of control area strategy for relative bias and statistical efficiency across a range of scenarios.

⁴⁴ Figures for predictive risk score are 44.6% in strategy 3, compared with 58.0% in strategy 1. Full data are available on request.

Table 5: Standardised differences (%), before and after matching, in the study of the rapid response service, under various strategies for selecting the control population (all individual baseline variables entered into the genetic matching algorithm)

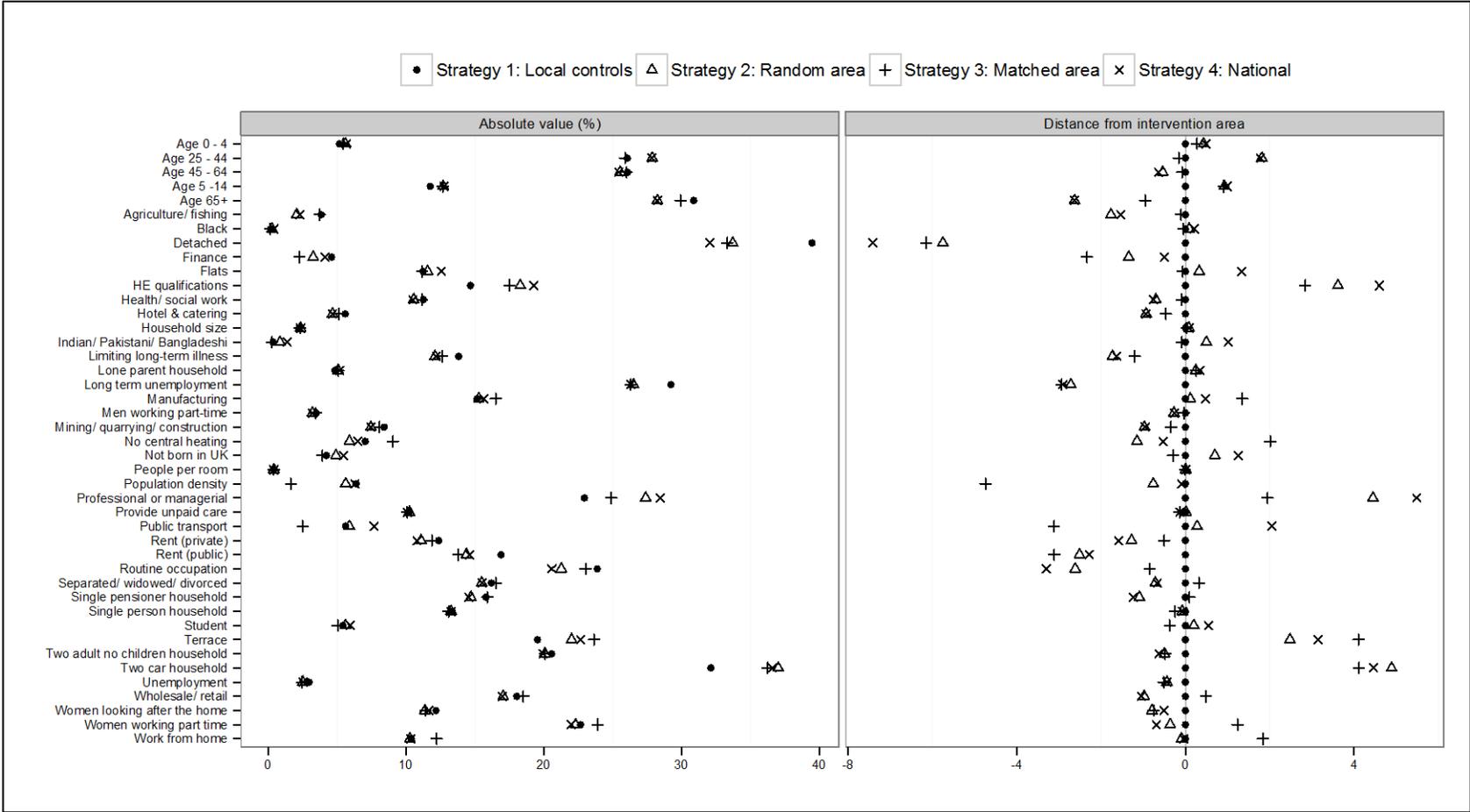
	Strategy 1: Local controls (392 potential controls)		Strategy 2: Random areas (500 potential controls) ⁴⁵		Strategy 3: Matched area (500 potential controls)		Strategy 4: National (16,000 potential controls)	
	Before	After	Before	After	Before	After	Before	After
Mean predictive risk score	59.6	4.7	55.9	4.5	55.5	4.7	56.6	1.3
Mean age	56.9	4.9	51.7	5.4	49.2	4.8	53.3	2.7
Mean number of emergency hospital admissions ⁴⁶	48.5	3.3	42.5	1.6	37.5	0.0	43.2	0.0
Female gender	22.5	-6.0	24.8	0.0	30.4	-2.0	26.1	0.0
Mean socioeconomic deprivation score	-1.5	1.3	21.8	7.5	13.8	6.6	20.1	0.7
Cancer prevalence	-3.2	2.6	5.2	2.6	2.6	7.8	2.5	0.0
Diabetes prevalence	3.5	5.7	11.3	0.0	6.2	2.8	9.4	0.0
Congestive heart failure prevalence	13.0	0.0	11.0	3.0	8.5	3.0	11.3	0.0
Ischemic heart disease prevalence	11.5	-12.1	12.8	0.0	16.5	4.8	13.4	0.0
Mean number of chronic conditions	26.0	-4.6	25.3	0.6	24.1	3.5	24.1	1.2
Mean number of planned admissions	13.2	9.0	14.6	6.5	11.9	3.9	14.6	0.0
Mean (absolute) standardised difference	23.6	4.9	25.9	5.4	23.3	4.0	25.0	0.5

Note: Negative values imply that the variable was lower on average in the intervention than matched control group.

⁴⁵ For reasons of space, standardised differences for strategy 2 are the medians over all 32 possible geographies. However, there was substantial variation depending on the choice of geography. Age, for example, showed standardised differences that ranged from 38.3% to 66.6% before matching, depending on which area was chosen, and from -12.5% to 20.0% after matching. Ranges for the predictive risk score were from 46.7% to 61.3% before matching, and from 0.4% to 12.5% after matching. For the number of emergency hospital admissions, ranges were from 34.7% to 48.3% before matching, and from -9.9% to 11.5% after matching.

⁴⁶ Admission counts are over the year prior to enrolment.

Figure 5: Area-level variables in the study of the rapid response service, under various strategies for selecting the control population



Notes: The first panel shows the values of each of the 43 area-level variables under each strategy, while the second panel shows relative differences from the intervention area. Figures for Strategy 2 ('random area') are medians over all 32 possible geographies. Figures for Strategy 4 ('national') are weighted means over the 32 geographies (weighted for population size).

Table 6: Estimated treatment effects of the rapid response service on emergency hospital admissions, under various strategies for selecting the control population

	Strategy 1: Local controls	Strategy 2: Random areas	Strategy 3: Matched area	Strategy 4: National
Relative risk (95% confidence interval)				
With all baseline variables included in the genetic matching	2.07 (1.46 to 2.94)	Median: 2.23 (1.52 to 3.28) Minimum: 1.35 (1.00 to 1.82) Maximum: 3.87 (2.43 to 6.16)	1.93 (1.37 to 2.73)	2.07 (1.47 to 2.92)
With age and predictive risk score omitted from the genetic matching	2.76 (1.89 to 4.03)	Median: 2.42 (1.61 to 3.63) Minimum: 1.66 (1.21 to 2.27) Maximum: 3.87 (2.32 to 6.44)	2.32 (1.57 to 3.42)	2.23 (1.55 to 3.21)
Absolute risk difference (95% confidence interval)				
With all baseline variables included in the genetic matching	0.28 (0.27 to 0.29)	Median: 0.30 (0.28 to 0.31) Minimum: 0.14 (0.13 to 0.15) Maximum: 0.40 (0.39 to 0.41)	0.26 (0.25 to 0.27)	0.28 (0.27 to 0.29)
With age and predictive risk score omitted from the genetic matching	0.34 (0.33 to 0.35)	Median: 0.31 (0.30 to 0.33) Minimum: 0.21 (0.20 to 0.23) Maximum: 0.40 (0.39 to 0.41)	0.31 (0.29 to 0.32)	0.30 (0.28 to 0.31)

4.6 Design of the simulation study

The simulations generalised previous examples (Drake 1993) to allow for observed and unobserved confounding at both individual and area levels (Table 7). One individual-level variable ($x_{1,1}$) was observed, while another ($x_{2,1}$) was unobserved. Two area-level variables ($x_{1,2}$ and $x_{2,2}$) were observed, while a third one ($x_{3,2}$) was unobserved. We considered this to be the minimum number of covariates needed to test a sufficiently wide range of scenarios.

Table 7: Design of the simulation study, and associations assumed in the base case scenario

	Level	Observation	Structure	Strength of relationship with intervention assignment (α)	Strength of relationship with outcome (β)
$x_{1,1}$	Individual	Observed	$x_{1,1}$ and $x_{2,1}$ are correlated	0.5	0.30
$x_{2,1}$	Individual	Unobserved		Range 0.1-0.3	0.15
$x_{1,2}$	Area	Observed	Independent	n/a	0.01
$x_{2,2}$	Area	Observed	Determines the mean of $x_{2,1}$	n/a	0.05
$x_{3,2}$	Area	Unobserved	Independent	n/a	0.06

In designing the simulations, we were mindful that researchers often have access to data at the aggregate level, but not to their individual-level counterparts. In the case study, for example, we had access to socioeconomic deprivation scores defined at a small area level, but not to the person-level equivalents. Similarly, in the sensitivity analysis that omitted age from the genetic matching algorithm, a control area was still matched to the intervention area on the overall age distribution. To mimic this situation, we assumed that the mean of the unobserved individual-level variable ($x_{2,1}$) differed between areas, but that this mean value corresponded to the value of the observed area-level variable $x_{2,2}$. Thus, control areas could be selected to minimise differences in the distribution of $x_{2,1}$ between the intervention and control area.

4.6.1 Generating baseline data

The three area-level variables ($x_{1,2}$ - $x_{3,2}$) were generated for each of 49 geographic areas by sampling from independent standard normal distributions (*i.e.*, mean 0, variance 1, pairwise correlations 0). Additionally, we assumed values of $x_{1,2}=x_{2,2}=x_{3,2}=1$ in the intervention area. This meant that, under the response models that are described below, the intervention area had atypical outcomes under control, as seems reasonable for areas that intervene to affect these outcomes. The assumption

also meant that the intervention area had an atypical distribution of the unobserved individual-level covariate, $x_{2,1}$ (*i.e.*, mean 1 as opposed to the expected level of 0), as would in general be the case.

The individual-level variables $x_{1,1}$ and $x_{2,1}$ were then generated for 1,000 individuals in each of the 50 areas, by sampling from a bivariate normal distribution, with means of 1 and $x_{2,2}$, respectively, a common variance of 1, and correlation 0.2.

Following previous simulation studies (Drake 1993; Austin, Grootendorst, and Anderson 2007), we assumed that the probability of receiving the intervention (the true propensity score) was a logistic function of the individual-level variables:

$$\Pr(t = 1 | x_{1,1}, x_{2,1}) = [1 + \exp\{-\alpha_0 + \alpha_{1,1}x_{1,1} + \alpha_{2,1}x_{2,1}\}]^{-1}$$

Here, $\alpha_{1,1}$ and $\alpha_{2,1}$ controlled how predictive the two individual-level variables were of intervention assignment, and therefore one aspect of confounding. The final coefficient (α_0) could be calibrated so that, in expectation across repeated simulations, a given proportion ($N\%$) of the individuals in the intervention area would receive the intervention.

4.6.2 Forming matched control groups

After baseline data had been generated, the different strategies to select control populations were applied. The matched control area for strategy 3 was selected as the one that minimised the Euclidean distance from the intervention area with respect to the two observed area-level variables ($x_{1,2}$ and $x_{2,2}$). Although the Mahalanobis metric could have been used (Rosenbaum and Rubin 1985), the result would have been the same as $x_{1,2}$ and $x_{2,2}$ were independent by assumption. The unobserved area-level variable $x_{3,2}$ could not be taken into account when selecting the matched control area. The individual-level variables were also not taken into account, on the assumption that, in health services and outcomes research, individual baseline data are typically not collected until after the study areas have been chosen. The control area for strategy 2 (convenience sample) was selected at random, while strategy 4 pooled potential controls from across the 49 non-intervention areas.

Matched control groups were formed at the individual level under each of the strategies. This was done before outcome data were simulated, so that matching was blind to outcome (Rubin 2008). The relative simplicity of the simulation design (in particular, with regard to the number of covariates) meant that, unlike in the case study, genetic matching was not required to balance the observed characteristics. Instead, the propensity score was estimated by applying logistic regression to data from the intervention area, and matches were formed using nearest neighbour matching on this estimated propensity score (1-1, with replacement). Since $x_{2,1}$ was unobserved, it was omitted from the propensity score model, and the model contained only a single variable, $x_{1,1}$ (more generally, this

variable might represent a weighted vector of many variables). When using external controls, the coefficients from the local propensity score model were applied to individuals in the control area.⁴⁷

Balance on the individual-level baseline variables was assessed under each strategy by reporting standardised differences. We also report mean values of the area-level variables.

4.6.3 *Generating outcomes and assessing treatment effects*

A dichotomous outcome was simulated according to a response model used in previous simulation studies (Drake 1993):

$$\begin{aligned} & Pr(Y = 1 | x_{1,1}, x_{2,1}, x_{1,2}, x_{2,2}, x_{3,2}, t) \\ & = [1 + \exp\{-(\beta_0 + \beta_{1,1}x_{1,1} + \beta_{2,1}x_{2,1} + \beta_{1,2}x_{1,2} + \beta_{2,2}x_{2,2} + \beta_{3,2}x_{3,2} + \delta t)\}]^{-1} \end{aligned}$$

The coefficients labelled $\beta_{i,j}$ determined how predictive the baseline variables were of the outcome. The binary variable, t , indicated whether the individual received the intervention ($t=1$) or not ($t=0$). The true intervention effect was denoted by δ and assumed to be zero.

Intervention effects were estimated under each strategy using the absolute risk difference (difference in proportions), relative to the corresponding matched control group. Bias was calculated by comparing the estimated treatment effect with the true intervention effect (*i.e.*, zero). We report mean bias over 20,000 simulations. We also obtained the Mean-Squared Error (MSE) by squaring the difference between the estimated and true intervention effects and averaging over all simulations.

4.6.4 *Calibration of the simulation study*

The simulations were calibrated to the HES data by taking emergency hospital admission to be the outcome, as in the case study, and then conducting sensitivity analysis for the response model and for the true propensity score model. In calibrating the level of individual confounding under the base case scenario, we assumed that the unobserved confounder was less predictive of intervention assignment and outcome than the observed confounder, and adopted $\alpha_{1,1}=0.5$, $\alpha_{2,1}=0.2$, $\beta_{1,1}=0.3$ and $\beta_{2,1}=0.15$. See Appendix A for a derivation of these values using national HES data (page 99).

An important aspect of the simulation design concerned the amount of explained and unexplained variation in outcomes between areas (this variation being generated through $x_{1,2}$, $x_{2,2}$ and $x_{3,2}$). To calibrate this aspect of the simulations, we assessed to what extent the risk of emergency hospital admission varies between similar individuals living in different areas of England, again using national HES data. Such between-area variation was assessed using the Median Odds Ratio (MOR) (Larsen and Merlo 2005), which is defined as the odds ratio that would be expected, in median, between

⁴⁷ Although the propensity score model could have been fitted using potential controls from external areas, in this instance, the result would have been the same, as there was only a single variable in the empirical propensity score model.

people with the same individual-level variables selected from two randomly-chosen areas. The MOR was calculated as 1.08 for people aged 70 or over in England - see Appendix A (page 99).

We made a conservative assumption about the amount of area-level variation that was explained by the observed area-level variables. Thus, we calibrated the simulation to two specific area-level variables, namely overall socioeconomic deprivation score and overall hospital admission rate (leading to $\beta_{1,2}=0.01$ and $\beta_{2,2}=0.05$). The remainder of the variation was assumed to be unexplained. A preliminary simulation showed that setting $\beta_{3,2}$ to approximately 0.06 gave an MOR of 1.08, and that approximately 70% of the resulting variation was unexplained (*i.e.*, due to the unobserved variable).⁴⁸

We calibrated the intercept of the true propensity model (α_0) to give an intervention saturation ($N\%$) of 30%, as in the case study.

4.6.5 Scenarios tested

We compared the bias and MSE resulting from each of the strategies under the following scenarios for the response model and for the true propensity score model.

Scenarios for the response model were:

1. As described above, the 'base case' scenario was calibrated to the associations seen for emergency hospital admissions in HES data. Thus, $\beta_{1,1} = 0.3$, $\beta_{2,1}=0.15$, $\beta_{1,2}=0.01$, $\beta_{2,2}=0.05$, $\beta_{3,2}=0.06$, and MOR=1.08.
2. The 'simple confounding' scenario assumed no confounding except through the observed individual-level variable ($\beta_{2,1}=\beta_{1,2}=\beta_{2,2}=\beta_{3,2}=0$, $\beta_{1,1}=0.3$). This was the ideal situation, under which all of the evaluation designs were expected to perform well.
3. The 'no area-level variation' scenario assumed no systematic variation in outcomes between areas, other than through the individual-level variables ($\beta_{1,2}=\beta_{2,2}=\beta_{3,2}=0$). Thus, the MOR was 1.
4. The 'no unexplained area-level variation' scenario assumed that all variation in outcomes between areas could be explained ($\beta_{1,2}=0.01$, $\beta_{2,2}=0.05$, $\beta_{3,2}=0$). Individual-level confounding was the same as in the base case.
5. The 'high unexplained area-level variation' scenario targeted a higher MOR of 1.3, which required $\beta_{3,2} = 0.3$. Other coefficients were the same as in the base case, implying that 95% of the area-level variation was unexplained.

⁴⁸ To calculate the percentage of MOR that was due to observed variables, we calculated MOR when $\beta_{3,2}=0$, and expressed this MOR as a proportion of total MOR.

Scenarios for the true propensity score model were:

- The central assumption, that $\alpha_{1,1}=0.5$, $\alpha_{2,1}=0.2$ and $N\% =30\%$.
- Lower and higher saturation ($N\%$), of 10% and 50%.
- Lower and higher confounding through the unobserved individual-level variable (*i.e.*, $\alpha_{2,1}$ equal to 0.1 and 0.3).

Finally, we repeated the simulations with a normally distributed, rather than dichotomous, outcome, and when matching without replacement, rather than with replacement.

4.7 Results of the simulation study

4.7.1 Standardised differences

The matching algorithm was generally able to find matched control groups that were closely balanced on the observed individual-level variable, regardless of which strategy was used to define the control population. For example, in strategy 1 (local controls), the matched control group had a mean of 1.352 on the observed variable, versus 1.354 for the intervention group in the base case scenario, leading to a standardised difference of 0.23% (Table 8). Increasing the saturation increased the standardised difference under strategy 1, as the supply of potential control patients available from within the local area became more limited. The standardised difference also increased under this strategy when the unobserved person-level variable became more predictive of intervention status. This led to greater before-matching differences on the observed variable, because of the correlation assumed between the individual-level variables. Although standardised differences were low under strategy 1, selecting controls from other areas could reduce them still further. Standardised differences were no more than 0.13% under strategies 2 and 3, and less than 0.01% under strategy 4.

Each of the approaches for selecting the control population led to large imbalances on the unobserved individual-level variable ($x_{2,1}$), especially when this variable was strongly predictive of intervention status (Table 9). Strategy 1 (local controls) produced a standardised difference of 19.40% under the base case scenario, whereas using strategy 3 (matched control area) resulted in a smaller standardised difference, of 18.14%. The relative advantage of strategy 3 over strategy 1 increased with higher saturation and stronger confounding; at low saturation levels (10%), strategy 1 produced the lower standardised differences. Using controls from random areas or from a national sample produced very large standardised differences on the unobserved variable across all scenarios.

While strategy 1 (local controls) exactly balanced the three area-level variables ($x_{1,2}$, $x_{2,2}$ and $x_{3,2}$) strategy 3 could only balance the two observed area-level variables in expectation (mean 1.0 and standard deviation 0.2, compared with a value of 1.0 in the local area). Strategy 3 could not balance

the unobserved area-level variable (mean 0, standard deviation 1). Strategies 2 and 4 led to large imbalances on all area-level variables.

Table 8: Balance in the simulation study, under various strategies for selecting the control population (observed individual-level confounder)

N%	$\alpha_{2,1}$	Means (standard deviations) after matching					Standardised differences (%) after matching			
		Treated	Matched controls from strategy				Strategy			
			1	2	3	4	1	2	3	4
10%	0.1	1.455 (0.099)	1.453 (0.099)	1.454 (0.099)	1.454 (0.099)	1.455 (0.099)	0.18	0.12	0.12	0.00
	0.2	1.471 (0.100)	1.469 (0.099)	1.470 (0.099)	1.470 (0.099)	1.471 (0.100)	0.19	0.13	0.13	0.00
	0.3	1.486 (0.098)	1.484 (0.098)	1.484 (0.098)	1.484 (0.098)	1.486 (0.098)	0.20	0.13	0.13	0.00
30%	0.1	1.344 (0.056)	1.342 (0.056)	1.343 (0.056)	1.343 (0.056)	1.344 (0.056)	0.22	0.07	0.07	0.00
	0.2	1.354 (0.056)	1.352 (0.056)	1.353 (0.056)	1.354 (0.056)	1.354 (0.056)	0.23	0.08	0.07	0.00
	0.3	1.363 (0.056)	1.361 (0.056)	1.362 (0.056)	1.362 (0.056)	1.363 (0.056)	0.24	0.08	0.07	0.00
50%	0.1	1.244 (0.044)	1.241 (0.043)	1.243 (0.043)	1.243 (0.043)	1.244 (0.044)	0.30	0.05	0.05	0.00
	0.2	1.250 (0.043)	1.247 (0.043)	1.250 (0.043)	1.250 (0.043)	1.250 (0.043)	0.31	0.05	0.05	0.00
	0.3	1.256 (0.043)	1.253 (0.043)	1.255 (0.043)	1.255 (0.043)	1.256 (0.043)	0.33	0.05	0.05	0.00

Note: the shaded section shows the base case scenario.

Table 9: Balance in the simulation study, under various strategies for selecting the control population (unobserved individual-level confounder)

N%	$\alpha_{2,1}$	Means (standard deviations) after matching				Standardised differences (%) after matching				
		Treated	Matched controls from strategy				Strategy			
			1	2	3	4	1	2	3	4
10%	0.1	1.176 (0.101)	1.080 (0.110)	0.097 (1.004)	1.040 (0.251)	0.090 (0.201)	9.66	109.03	13.76	109.64
	0.2	1.261 (0.100)	1.071 (0.109)	0.094 (1.003)	1.046 (0.251)	0.094 (0.202)	19.30	118.23	21.81	118.31
	0.3	1.345 (0.100)	1.062 (0.109)	0.093 (1.010)	1.047 (0.251)	0.100 (0.201)	28.74	127.31	30.27	126.61
30%	0.1	1.132 (0.058)	1.036 (0.076)	0.071 (0.998)	1.015 (0.240)	0.068 (0.164)	9.63	106.74	11.75	107.03
	0.2	1.198 (0.058)	1.006 (0.076)	0.072 (1.005)	1.018 (0.238)	0.071 (0.165)	19.40	113.90	18.14	113.99
	0.3	1.260 (0.057)	0.976 (0.077)	0.071 (0.996)	1.022 (0.237)	0.072 (0.165)	28.91	121.01	24.26	120.93
50%	0.1	1.094 (0.045)	0.998 (0.075)	0.040 (1.005)	1.001 (0.238)	0.048 (0.157)	9.65	106.03	9.29	105.21
	0.2	1.139 (0.044)	0.948 (0.075)	0.062 (1.003)	0.998 (0.235)	0.050 (0.158)	19.36	108.91	14.31	110.20
	0.3	1.183 (0.044)	0.900 (0.075)	0.052 (0.997)	0.998 (0.238)	0.051 (0.156)	28.83	115.30	18.93	115.41

Note: the shaded section shows the base case scenario.

4.7.2 *Bias and mean-squared error*

As would be expected, all strategies produced near-unbiased treatment effect estimates in the 'simple confounding' scenario, when the only confounder was the observed individual-level variable (Figure 6, first panel, and Table 10). When there was unobserved confounding at the individual level, but no area-level variation in outcomes (corresponding to an MOR of 1), using a matched control area still gave the least biased and most precise estimates (Figure 6, second panel, and Table 10). Similarly, when area-level variation in outcomes existed but was entirely explained by the observed variables, using a matched control area again produced the least biased estimates, though no longer the lowest MSE (Figure 6, third panel). The final two scenarios shown in Figure 6 include unexplained area-level variation in outcomes.

The base case scenario (Figure 6, fourth panel) had an MOR of 1.08, with 70% of this variation being unexplained. In this scenario, local controls gave the least biased and most precise estimates, with a bias of 0.24%. Strategies 2 and 4 were very biased, whereas using a matched control area produced a bias closer to the local approach (0.79%, see Table 10). The scenario with higher unexplained area-level variation (MOR=1.30, 95% unexplained), exaggerated the differences between the strategies still further (Figure 6, last panel).

Strategies 2 and 4 gave large but similar biases in all except for the 'simple confounding' scenario, because neither strategy addressed the possibility that the unobserved individual-level variable might be distributed differently within the intervention area to without, and thus there were large imbalances on that variable under both strategies (Table 9). Estimates from strategy 4 were more precise than those from strategy 2 (Table 10).

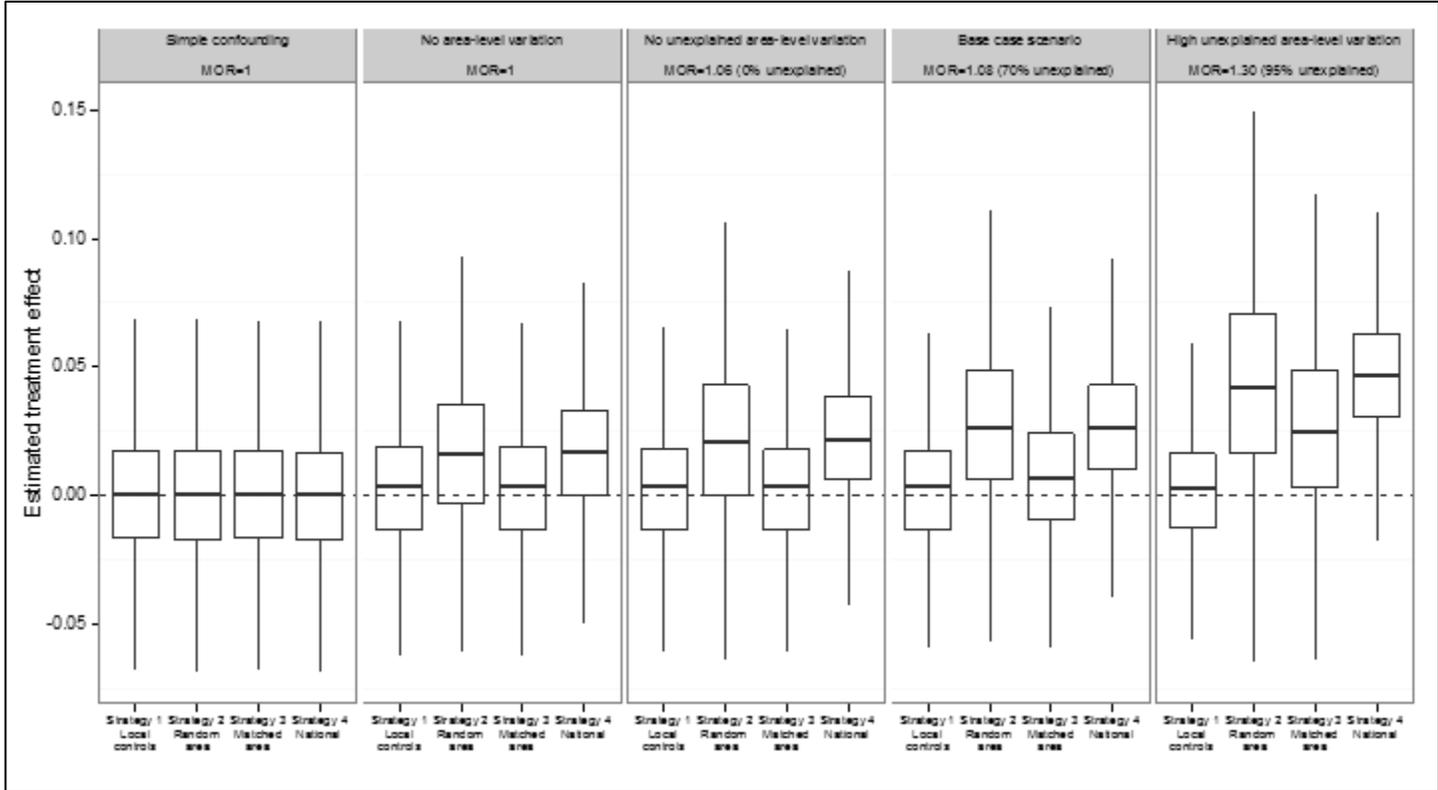
Our conclusions for the base case scenario did not change when varying the amount of confounding through the unobserved individual-level variable ($\alpha_{2,1}$). Both strategies 1 and 3 reported higher levels of bias at higher values of this parameter ($\alpha_{2,1}=0.3$, bias 0.31% and 0.85%, respectively), and lower biases at lower values ($\alpha_{2,1}=0.1$, bias 0.14% and 0.71%, respectively). Matching without replacement marginally increased the standardised differences obtained for the observed individual-level variable when using local controls, but the impact on the overall bias was very small in the base case scenario (see Table 12 and Figure 7, Appendix B, page 103). Using a normally distributed outcome gave a similar pattern to Figure 6 (see Figure 8, Appendix B, page 104).

Table 10: Bias (mean-squared error) in the simulation study, under various strategies for selecting the control population

	Strategy 1: Local controls	Strategy 2: Random areas	Strategy 3: Matched area	Strategy 4: National
Central assumptions ($N = 30\%$)				
Simple confounding	0.04 (6.51)	-0.02 (6.71)	0.02 (6.67)	-0.02 (6.50)
No area-level variation	0.26 (5.92)	1.64 (11.18)	0.24 (5.90)	1.68 (8.97)
No unexplained area-level variation	0.26 (5.61)	2.23 (14.97)	0.23 (5.71)	2.18 (10.89)
Base case scenario	0.24 (5.36)	2.73 (17.69)	0.79 (6.53)	2.66 (13.14)
High unexplained area-level variation	0.17 (4.49)	4.74 (41.39)	2.76 (19.69)	4.68 (27.79)
Sensitivity analysis: low saturation ($N = 10\%$)				
Simple confounding	-0.02 (19.49)	-0.05 (19.36)	0.01 (19.48)	0.03 (19.72)
No area-level variation	0.26 (16.86)	1.60 (22.76)	0.25 (17.26)	1.66 (21.10)
No unexplained area-level variation	0.25 (16.74)	2.19 (26.26)	0.29 (16.25)	2.18 (22.66)
Base case scenario	0.22 (15.53)	2.70 (29.17)	0.78 (17.28)	2.65 (24.64)
High unexplained area-level variation	0.20 (12.54)	4.59 (49.36)	2.69 (28.80)	4.56 (37.29)
Sensitivity analysis: high saturation ($N = 50\%$)				
Simple confounding	0.02 (4.06)	0.00 (4.07)	-0.01 (4.04)	-0.03 (4.06)
No area-level variation	0.26 (3.75)	1.66 (8.81)	0.20 (3.80)	1.65 (6.57)
No unexplained area-level variation	0.24 (3.51)	2.25 (13.06)	0.24 (3.61)	2.18 (8.59)
Base case scenario	0.23 (3.36)	2.74 (15.80)	0.74 (4.31)	2.70 (11.05)
High unexplained area-level variation	0.20 (2.73)	4.89 (41.92)	2.79 (18.49)	4.78 (26.57)

Notes: the shaded section shows the base case scenario. $\alpha_{2,1} = 0.2$ throughout. Bias and mean-squared error are expressed as a percentage of the population size.

Figure 6: Box plots of the estimated treatment effects in the simulation study, under various strategies for selecting the control population



Notes: Based on 20,000 replications from the simulation experiment. The horizontal dashed line represents the true treatment effect. Saturation = 30% and $\alpha_{2,1} = 0.2$ throughout.

4.8 Discussion

Careful design is of paramount importance to observational studies since, however advanced the analytical method, the study is likely to be biased if the underlying assumptions are not met (Rubin 2008). Investigators have used a range of approaches to define the control population when evaluating health care interventions, but the relative benefits of some popular design choices (in particular, local or external control populations) have rarely been directly assessed (Rosenbaum 1987; Stuart and Rubin 2008). The findings of the case study and simulations can assist investigators in deciding on their strategy for control area selection.

In the case study, balance on individual-level variables was improved by using controls from a matched area rather than locally. When we induced unobserved confounding by omitting two prognostic variables from the matching algorithm (namely, age and predictive risk score), balance on these unobserved variables was also better when selecting controls from a matched area rather than locally. The simulations built upon the case study, and identified two criteria that were necessary for matched control areas to produce the best balance on individual-level variables. First, intervention saturation had to be relatively high - at least 30%, the level seen in the case study. Second, the relationship between the unobserved variable and treatment assignment had to be relatively strong, as was the case for age and predictive risk score in the case study. The intuition behind the second condition is that, if the relationship between those variables was weak, then relatively good balance can be achieved locally. Meanwhile, selecting controls from outside of the intervention area risks systematic differences in the distribution of the unobserved individual-level variable or area-level variation in the outcome.

In the case study, treatment effects were more robust to induced, unobserved confounding when using a matched control area than local controls. The matched control area in the case study was selected using an established set of variables. We were reassured that, given the wide range of treatment effects that could be produced when using external areas (see Column 3, Table 6), treatment effects produced using the matched area were similar to those estimated locally. However, it was unclear which strategy was least biased. The case study could not assess bias, which can arise from differences at the area level as well as at the individual level.

The simulations showed that a matched control area produced the lowest bias of all the strategies, provided that, first, it produces better balance at the individual level and, second, area-level variation either does not exist or can be largely explained by the observed area-level variables. In the terminology of Section 4.4, this means that error terms 1 and 2 are minimised while error term 4 is close to zero. In other scenarios, where there was substantial unexplained variation in outcomes, a research design using a

matched control area was more biased than one using local controls. In other words, the increases in error term 4 associated with moving from local to external controls outweighed reductions in terms 1 and 2. For hospital admissions (represented by the base case scenario in Table 10), significant unexplained variation was likely, and so local controls gave least bias. Previous research has found that regional variation in hospital admission rates is partly due to differences in service design, admission thresholds and culture (Joynt and Jha 2012) - factors that are not often captured in routinely collected data. Had we better data on area-level confounders, we would have been closer to the 'no unexplained variation' scenario than the 'base case' scenario, and thus might have preferred to use a matched control area.

Translating the results of the simulation back into the case study, we infer that the preferred estimate of the treatment effect relies on local not external controls. Thus, our preferred estimate of the relative risk of emergency admissions is the local estimate, *i.e.* 2.07 (95% CI, 1.46 to 2.94). This information does not negate the value of using multiple control groups, as proposed by Campbell (1969) and (Rosenbaum 1987). Indeed, under strategy 2, we repeated the matching algorithm 32 times, once for each potential choice of control population. Every analysis reported more emergency hospital admissions among intervention than control patients, increasing the degree of confidence we place in this finding. However, precise effect sizes varied greatly depending on the choice of area, from a rate ratio of 1.35 to one of 3.87. The considerations described above lead us to prefer an estimate of around double. This estimate is likely to be least affected by unobserved confounding, although it is still susceptible to it.

4.8.1 Limitations and future research

Although we considered a range of scenarios for both the response model and the propensity score model, the simulations were still limited in some respects. We assumed that the areas had the same population size whereas, if some areas were larger than others, then this would increase their attractiveness as sources of controls, all other factors being equal. We also assumed that the distribution of the unobserved individual-level variable $x_{1,2}$ differed between areas in ways that could be controlled for by careful selection of the control area. This reflected a common situation in which individual-level variables are manifested at area levels. For example, although data on individuals' education levels might not be available, estimates might be obtained for the average education level of residents of different areas, for example from surveys. If such information is not available, then using a matched control area will generally not balance unobserved variables, as is apparent from the results of strategy 2 in Table 9. We also assumed that individual-level variables could not be used to select the matched control area. This reflects a common situation in which data can only be obtained from a small number of areas, either because of the cost of data collection or because of information governance considerations. However, in

other situations, national individual-level data may be available from administrative systems (Steventon et al. 2012b), and these could be used to select control areas.

We modelled the relatively simple situation in which there is a single intervention unit, and assumed that this was prone to atypical levels of the outcome under control, as would generally be the case. One could apply the same strategies 1-4 if there were several intervention units, but in general one would select several matched control areas under strategy 3. These areas could be selected by first constructing a propensity score that models the decision of each area to offer the intervention. Each intervention area would then be matched to a control area (for example, using nearest-neighbour matching on the propensity score or genetic matching), and then an individual-level matching algorithm would be applied to the corresponding area pairs. Other strategies also become available in this more general setting. Griswold and Localio (2010) fitted a single, multilevel propensity score model using individual-level data from several hospitals. This aimed to model both area-level and individual-level decision-making and resulted in a one-step matching process. These authors could not directly assess bias and statistical efficiency in their setting, but future work could conduct simulations similar to those in the current study to assess the statistical properties of their approach. Abadie et al. (2010) considered the same situation but created synthetic control groups by weighting the outcomes of several control areas. In their example, this approach gave good balance on the historical trend in the outcome. Stuart and Rubin (2008) considered a simpler situation with one intervention area and one external area. Although their method did not deal with unobserved confounding at the individual level, they made an adjustment for area-level group differences that could be developed for the more general scenario.

This paper addressed approaches to selecting control areas when estimating sample average treatment effects. If the estimand of interest was a population average treatment effect, then one would need to deal with many of the same issues about confounding at individual and area levels (Hartman et al. 2015). In estimating sample average treatment effects, we used a relatively simple matching method (nearest-neighbour matching) and a relatively simple estimator (difference in proportions). We also addressed scenarios in which observed individual-level variables were easily balanced, as is apparent from Table 8. If balance on these variables were harder to achieve, then more sophisticated matching methods (Hansen 2008b; Sekhon and Grieve 2012) or more complex estimators (Bang and Robins 2005; Ho et al. 2007) may be helpful. However, these would not overcome unobserved confounding, and so we expect that our findings would hold true in the more general setting.

Variance estimation following matching was not the focus of the current paper but its importance for inference is recognised. The case study used an approach to estimating confidence intervals that reflected

the paired nature of the matched data set (Agresti and Min 2004). Further work could incorporate recent developments in matching methods to allow for the dependencies that arise in the data when matching with replacement, while also recognising any clustering of individuals within control areas (Abadie and Imbens 2012).

Finally, we assumed that a standardised data set exists across areas, and so we did not take into account the possibility that measurement varied between areas. Such variation would introduce additional biases and uncertainty when selecting external controls. On the other hand, we did not account for spillover effects, whereby the intervention affects the care received by local untreated individuals. These might lead investigators to prefer external controls.

4.8.2 Conclusions

Our findings underscore the importance of considering the data generating process underlying observational data sets, which is a multilevel phenomenon when patients come from more than one higher-level unit (*e.g.*, from several geographic areas). The theory behind propensity score matching (Rosenbaum and Rubin 1983), when applied to a multilevel context, emphasises the need to model correctly the process by which some areas offer the intervention and others do not, as well the process by which some individuals receive the intervention and others do not. This observation explains why large biases arose when using randomly chosen control areas or national controls. Investigators may trade off individual-level versus area-level confounding by selecting matched controls from a matched area rather than from within the intervention area. Box 4 summarises some factors to consider when selecting control populations for observational studies.

Although there have been advances in analytical methods, design issues tend to be relatively neglected in observational studies, and there is limited guidance to help researchers improve study design and assess whether a data set is adequate to answer the questions being asked of it (Rubin 2010). We have provided a set of considerations that relate to the control population, and append code to help investigators undertake similar simulations to those presented here at the design stage of future observational studies (page 228), if there is doubt about which approach is best. This could complement sensitivity analyses using multiple control groups.

Box 4: Factors to consider when selecting control populations

Local controls may be preferred when there is:

- Low or moderate intervention saturation; and
- Low risk of unobserved confounding at the individual level.

Controls from a matched area may be preferred when:

- High intervention saturation means there is a limited supply of controls from within the local area;
- Unobserved confounding is likely at the individual level, and the unobserved confounder is a relatively strong predictor of treatment assignment;
- The distribution of the unobserved confounder in the matched control area is likely to be similar to its distribution in the intervention area; and
- Area-level variation in outcomes either does not exist or can be largely explained by observed area-level variables that are accounted for in the matching.

Other considerations include the relative population sizes of the areas, spillover effects and differences in measurement.

4.9 Appendix A: Calibration of the simulation study

The simulations aimed to present a modest number of plausible scenarios to illustrate the main strengths and limitations of the study designs, rather than to be exhaustive in the range of scenarios covered.

Therefore, although many applied studies allow for a large number of confounders, the simulation design was stylised and contained only two at the individual level: $x_{1,1}$ (which was observed) and $x_{2,1}$ (which was unobserved). This appendix describes how the base case scenario was calibrated to national hospital data, in terms of these individual-level variables, the three area-level variables, and the intervention saturation. We took the outcome to be future emergency hospital admission, as in the case study.

4.9.1 Individual-level variables

As one of the strongest predictors of future emergency hospital admissions is the number of such admissions experienced in the past (Roland et al. 2005; Billings et al. 2006), for the purposes of calibration, $x_{1,1}$ was taken to be the number of admissions experienced in the 1-6 months before the intervention start date, while $x_{2,1}$ was taken to be the number of admissions experienced in the 7-12 months before. The scenario is, therefore, that the analysis is adjusted for the number of admissions occurring immediately before the intervention started, but not for the more distant history.

In the case study data, which concerned people aged 70 or over, the two individual-level variables were observed to have correlation of 0.18. Both variables were associated with intervention assignment, with logistic regression coefficients of 0.25 and -0.12, respectively, after normalisation. We assumed that the observed confounder was associated more strongly than this with treatment allocation (coefficient 0.5), as in practice several observed confounders would be adjusted for, and these would explain a greater proportion of the variation in intervention assignment than one variable would do by itself. In the simulations, we set $\alpha_{1,1}=0.5$, and took various values for $\alpha_{2,2}$ (0.1, 0.2 and 0.3).

Both variables were associated with the outcome at the 5% significance level, with coefficients 0.29 and 0.14, respectively, after normalisation. We took $\beta_{1,1}=0.3$ and $\beta_{2,1}=0.15$.

Area-level variation

An important aspect of the simulation design concerned the amount of explained and unexplained variation in outcomes between areas (this variation being generated through the three area-level variables, $x_{1,2}$, $x_{2,2}$ and $x_{3,2}$). To calibrate this aspect of the simulations, we assessed to what extent the risk of emergency hospital admission varies between similar individuals living in different areas of England, using national HES data. Rather than restrict our focus to the district councils considered in the case study

(which tend to be predominately rural), we considered all of the 152 health care administrative areas that existed in 2008 ('Primary Care Trusts'). We restricted our attention to patients who were aged 70 or over with at least one recent admission, as in the case study. Between-area variation in emergency hospital admissions in 2008 was assessed using the Median Odds Ratio (MOR) (Larsen and Merlo 2005), which is defined as the odds ratio that would be expected, in median, between people with the same individual-level variables selected from two randomly-chosen areas. The MOR was estimated using a random effects logistic model, adjusting for the individual-level variables that entered the predictive risk model used in the case study (Billings et al. 2006). The MOR was calculated as 1.08 for people aged 70 or over.

The area-level variables $x_{1,2}$ and $x_{2,2}$ were taken to be an area-level deprivation score and the overall hospital admission rate for the area, respectively. Both were associated with outcome (emergency hospital admission) at the 5% significance level, with coefficients 0.01 and 0.05, respectively, after normalisation.

The role of the unobserved area-level confounder ($x_{3,2}$) was to capture unexplained variation in hospital admission rates between areas. A preliminary simulation revealed that setting $\beta_{3,2} = 0.06$ would be expected to produce an MOR of 1.08 in the final simulated data set, given $\beta_{1,2}=0.01$ and $\beta_{2,2}=0.05$.

Intervention saturation

The final parameter used in the simulation study related to the proportion of residents in the intervention area who received the intervention. In the case study, a total of 491 people aged 70 or over with a history of hospital admissions were eventually enrolled into the rapid response service, representing less than 1% of the total number of such people in the area. However, saturations of 20% and higher were modelled as it was assumed that researcher would have further narrowed down the population of potential controls (for example, by disease group). Higher saturations were tested as they are encountered in health services and outcomes research, and they helped to test the robustness of a method based on selecting controls from within the intervention area.

4.10 Appendix B: Sensitivity analysis

This appendix contains sensitivity analysis when matching without replacement, and when using a normally distributed outcome.

4.10.1 *Matching without replacement*

The principal analysis presented in the main paper assumed that matching was conducted with replacement, so that a given individual could act as a matched control for more than one intervention patients. Matching without replacement marginally increased the standardised differences obtained for the observed individual-level variable when using local controls (see Table 11). It did not substantially impact standardised differences for the unobserved individual-level variable, or the overall bias (see Table 12 and Figure 7).

4.10.2 *Normally distributed outcome*

Figure 8 compares the box-and-whisker plots presented in the main paper with those produced using a normally distributed outcome. This had standard deviation 0.5 and mean equal to:

$$(\beta_0 + \beta_{1,1}x_{1,1} + \beta_{2,1}x_{2,1} + \beta_{1,2}x_{1,2} + \beta_{2,2}x_{2,2} + \beta_{3,2}x_{3,2} + \delta t)$$

The essential features of the original plot remains. As before, the 'base case' scenario is calibrated to the case study ($\beta_{1,2}=0.01, \beta_{2,2}=0.05, \beta_{3,2}=0.06, \text{MOR}=1.08$). The second scenario assumes there is no confounding except through the observed individual variable ($\beta_{2,1}=\beta_{1,2}=\beta_{2,2}=\beta_{3,2}=0$) while, in the third, there is no area-level confounding ($\beta_{1,2}=\beta_{2,2}=\beta_{3,2}=0$). The final scenario assumes higher unexplained variation in outcomes between areas ($\beta_{3,2}=0.3, \text{MOR}=1.3$). Saturation = 30% and $\alpha_{2,1} = 0.3$ throughout.

Table 11: Balance in the simulation study when matching with and without replacement

		Observed individual-level variable		Unobserved individual-level variable	
		With replacement	Without replacement	With replacement	Without replacement
Means (standard deviations)	Treated	1.363 (0.056)	1.363 (0.055)	1.260 (0.057)	1.259 (0.057)
	Strategy 1: Local controls	1.361 (0.056)	1.335 (0.049)	0.976 (0.077)	0.972 (0.057)
	Strategy 2: Random areas	1.362 (0.056)	1.361 (0.055)	0.071 (0.996)	0.063 (1.012)
	Strategy 3: Matched area	1.362 (0.056)	1.361 (0.055)	1.022 (0.237)	1.021 (0.235)
	Strategy 4: National	1.363 (0.056)	1.363 (0.055)	0.072 (0.165)	0.073 (0.163)
Standardised differences (%)	Strategy 1: Local controls	0.24	2.87	28.91	29.22
	Strategy 2: Random areas	0.08	0.20	121.01	121.86
	Strategy 3: Matched area	0.07	0.20	24.26	24.22
	Strategy 4: National	0.00	0.00	120.93	120.79

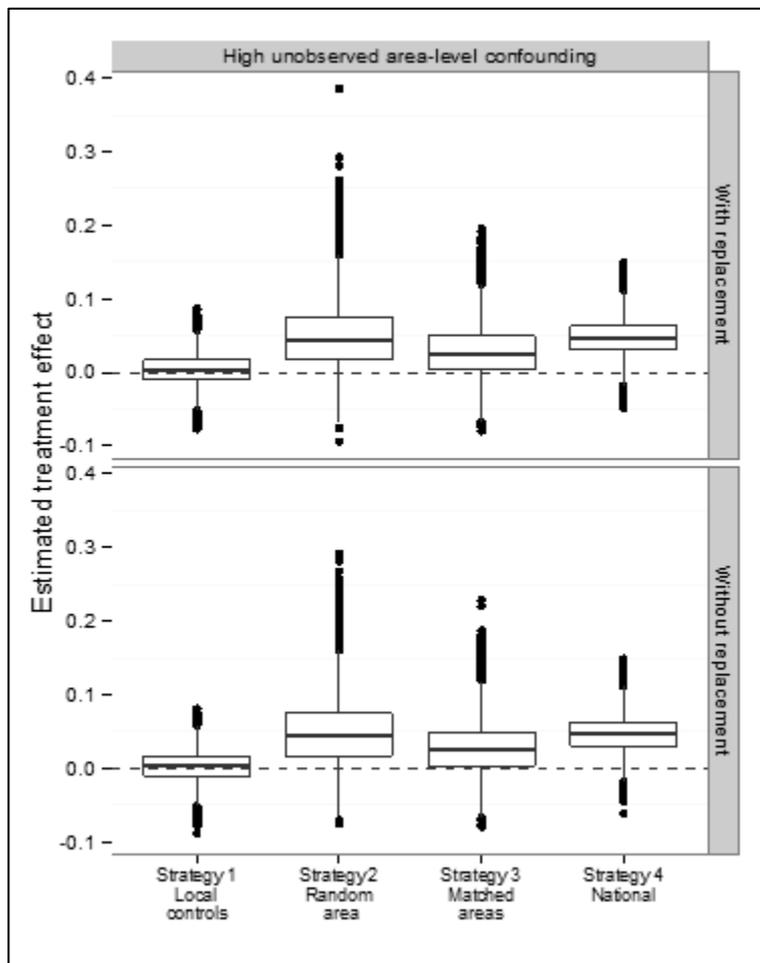
Note: base case scenario with 30% saturation and $\alpha_{2,1} = 0.3$.

Table 12: Bias (mean-squared error) in the simulation study when matching with and without replacement

	Strategy 1: Local controls	Strategy 2: Random areas	Strategy 3: Matched area	Strategy 4: National
With replacement	0.30 (4.40)	4.80 (42.29)	2.83 (20.27)	4.73 (28.28)
Without replacement	0.33 (4.41)	4.84 (42.67)	2.82 (19.86)	4.72 (28.09)

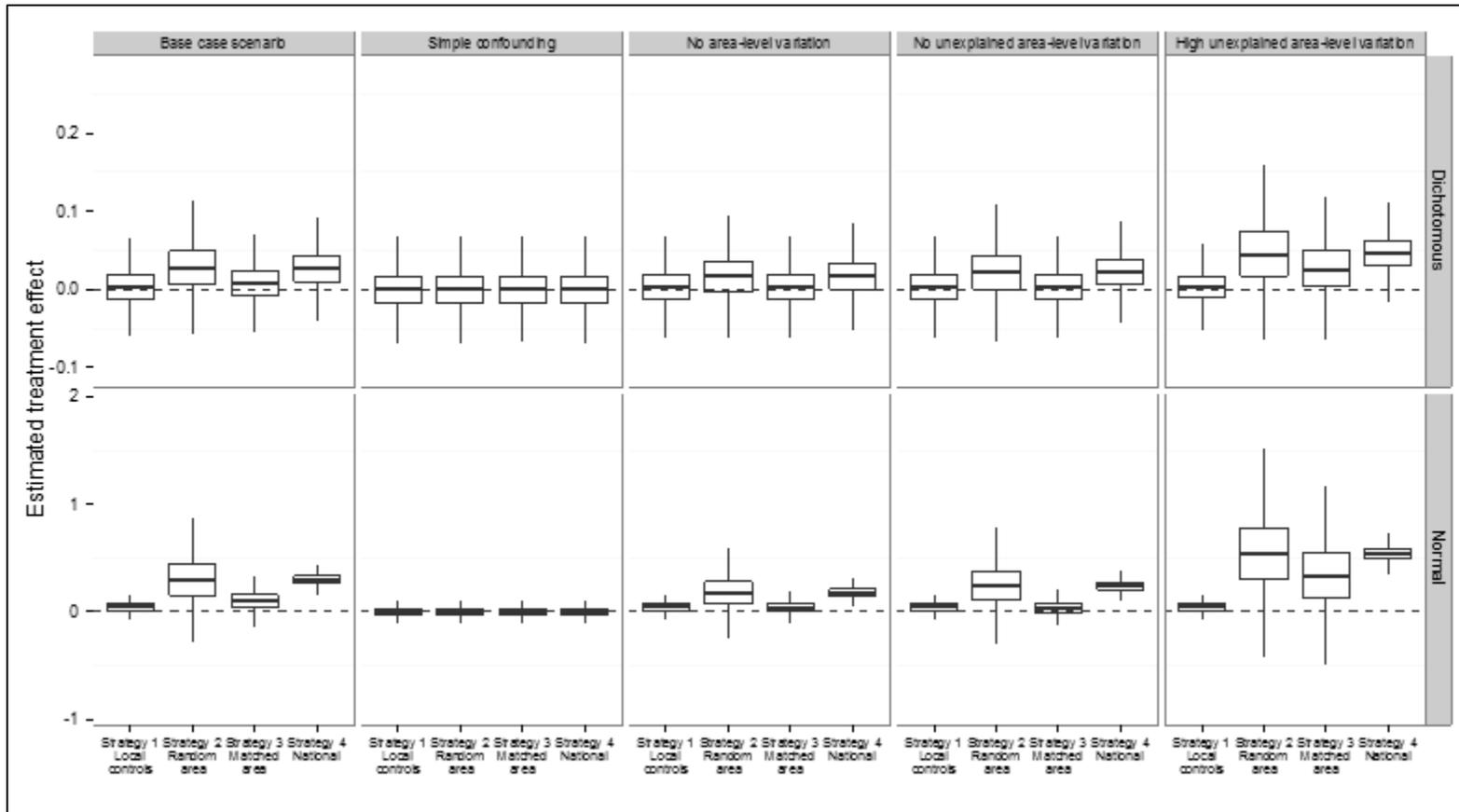
Note: high unobserved area-level confounding scenario with 30% saturation and $\alpha_{2,1} = 0.3$.

Figure 7: Box plots of the estimated treatment effects in the simulation study when matching with and without replacement



Notes: The horizontal dashed line represents the true treatment effect. 30% saturation and $\alpha_{2,1} = 0.3$.

Figure 8: Box plots of the estimated treatment effects in the simulation study when using dichotomous and normally distributed endpoints



The horizontal dashed line represents the true treatment effect. 30% saturation and $\alpha_{2,1} = 0.3$.

Chapter 5 Effect of telehealth on hospital use and mortality (research paper 3)

5.1 Preamble to research paper 3

This paper presents headline findings from the Whole Systems Demonstrator (WSD), a large cluster randomised trial that compared telehealth with usual care in a sample of more than 3,000 people from three areas of England. The WSD was funded by the Department of Health in response to concerns about the robustness of previous evaluative studies of telehealth, which had typically been small, of doubtful generalisability and of variable quality (Barlow et al. 2007b; Bergmo 2009).

As part of the evaluation, routine data were extracted from the electronic medical record of participating general practices, and linked at the person level to administrative data on hospital utilisation. This resulted in a data set of over one billion observations, which was used to assess outcomes and to calibrate a prognostic model of emergency hospital admissions. The estimated treatment effects were adjusted for between-group differences in the prognostic score. This was considered to be important since cluster randomised trials can be susceptible to selection bias.

The paper found that telehealth patients experienced fewer emergency hospital admissions and fewer deaths than controls during the 12-month follow-up period. However, the longitudinal administrative data revealed that emergency hospital admissions increased among control patients shortly after they were recruited into the trial. Thus, the paper raises concerns that the implementation of the trial protocol might have led some patients in the control group to seek additional care – a threat to generalisability that is examined using placebo tests in Chapter 6.

This paper has previously been published in the *BMJ* (Steventon et al. 2012a), under the Creative Commons Attribution Non Commercial (CC BY-NC 3.0) license. It has been reformatted for this thesis.

RESEARCH PAPER COVER SHEET

PLEASE NOTE THAT A COVER SHEET MUST BE COMPLETED FOR EACH RESEARCH PAPER INCLUDED IN A THESIS.

SECTION A – Student Details

Student	Adam Steventon
Principal Supervisor	Professor Richard Grieve
Thesis Title	Evaluating complex interventions using routinely collected data: Methods to improve the validity of randomised controlled trials and observational studies

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?	BMJ		
When was the work published?	2012		
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?*	No	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	Choose an item.

SECTION D – Multi-authored work

<p>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)</p>	<p>I contributed to the final protocol for the trial (Bower et al. 2011) and led the data collection and analysis for this paper. My specific roles included arranging for trial participants to be linked to HES data, and supervising a subcontractor (Bupa Health Dialog). I collaborated with the trial statistician (Dr Helen Doll) and representatives from the other themes in designing the details of the analytical approach, including the specification of the regression models and the presentation of the data. I conducted all analyses and produced the first draft of the manuscript, before incorporating comments from other co-authors and peer reviewers.</p>
---	---

Student Signature: _____

Date: _____

Supervisor Signature: _____

Date: _____

5.2 Abstract

Objective: To assess the effect of home-based telehealth interventions on the use of secondary health care and mortality.

Design: Pragmatic, multisite, cluster randomised trial comparing telehealth with usual care, using data from routine administrative datasets. General practice was the unit of randomisation. We allocated practices using a minimisation algorithm, and did analyses by intention to treat.

Setting: 179 general practices in three areas in England.

Participants: 3,230 people with diabetes, chronic obstructive pulmonary disease, or heart failure recruited from practices between May 2008 and November 2009.

Interventions: Telehealth involved remote exchange of data between patients and health care professionals as part of patients' diagnosis and management. Usual care reflected the range of services available in the trial sites, excluding telehealth.

Main outcome measure: Proportion of patients admitted to hospital during 12-month trial period.

Results: Patient characteristics were similar at baseline. Compared with controls, the intervention group had a lower admission proportion within 12-month follow-up (odds ratio 0.82, 95% confidence interval 0.70 to 0.97, $p=0.017$). Mortality at 12 months was also lower for intervention patients than for controls (4.6% *vs.* 8.3%; odds ratio 0.54, 0.39 to 0.75, $p<0.001$). These differences in admissions and mortality remained significant after adjustment. The mean number of emergency admissions per head also differed between groups (crude rates, intervention 0.54 *vs.* control 0.68); these changes were significant in unadjusted comparisons (incidence rate ratio 0.81, 0.65 to 1.00, $p=0.046$) and after adjusting for a predictive risk score, but not after adjusting for baseline characteristics. Length of hospital stay was shorter for intervention patients than for controls (mean bed days per head 4.87 *vs.* 5.68; geometric mean difference -0.64 days, -1.14 to -0.10 , $p=0.023$, which remained significant after adjustment). Observed differences in other forms of hospital use, including notional costs, were not significant in general. Differences in emergency admissions were greatest at the beginning of the trial, during which we observed a particularly large increase for the control group.

Conclusions: Telehealth is associated with lower mortality and emergency admission rates. The reasons for the short-term increases in admissions for the control group are not clear, but the trial recruitment processes could have had an effect.

5.3 Introduction

Efforts worldwide are dealing with the increasing prevalence of chronic disease among an ageing population. The past decade has seen the growing use of telehealth as one possible approach to this problem. Telehealth involves the remote exchange of data between a patient and health care professionals as part of the patient's diagnosis and health care management (McLean, Protti, and Sheikh 2011; Sood et al. 2007). Examples include the monitoring of blood pressure and blood glucose. Telehealth may help patients to better understand their health conditions by providing tools for self-monitoring, encourage better self management of health problems, and alert professional support if devices signal a problem. As a consequence, telehealth promises better quality and more appropriate care for each patient, as well as more efficient use of health care resources by reducing the need for expensive hospital care.

Some research suggests that telehealth can have a positive effect on patients with chronic disease, such as improved patient experiences, clinical indicators and quality of life, and reduced use of secondary health care (including emergency hospital admissions (McLean, Protti, and Sheikh 2011; Sood et al. 2007). Yet, other studies have found either no effect or a negative effect (Chaudhry, Mattera, and Krumholz 2011; Barlow et al. 2007b). Furthermore, such evidence is usually based on assimilating findings from a number of small trials, which could be difficult to generalise (Chaudhry, Mattera, and Krumholz 2011), and with many of these trials not meeting robust evaluation standards (Barlow et al. 2007b; Bergmo 2009). A recent review of self monitoring of blood glucose for people with diabetes concluded that there was a need for large controlled trials (Farmer et al. 2005).

Investment in telehealth has often been justified partly on the basis that its cost can be recovered by reductions in the use of secondary health care (Cruickshank et al. 2010). However, assessing the scale of such an effect is complicated. Simple study designs comparing stages before and after an intervention can produce misleading results by not having a control group to compare with, particularly if the patients selected for intervention have a history of emergency care. Such patients have a tendency to show reductions in use of emergency care over time (that is, regression to the mean) (Roland et al. 2005). Therefore, in the absence of a control group, whether observed reductions are the effect of the intervention is unclear.

Analyses of hospital use are further complicated by the fact that the distribution of admissions across patients can be highly skewed. Some high-risk patients account for a very high proportion of admissions (Billings et al. 2006). Therefore, small differences in the risk profile of patients receiving the intervention can greatly affect observed outcomes in terms of hospital admission. Several predictive risk models have been developed that use information from a person's health history to predict future hospital use (Nuffield Trust 2011; Wennberg et al. 2006), and can offer an opportunity for case mix adjustment. A further limitation on the size of previous evaluation studies has been the costs of obtaining information from patients, but it is now possible to extract information from operational administrative systems and use secure data linkage procedures to track resource use.

In 2006, the Department of Health in England published a white paper that included a focus on health and social care for people with long-term needs (Department of Health 2006). The strategy proposed a series of demonstrator pilots to drive whole systems redesign, supported by advanced assistive technologies. These technologies included telehealth, along with a system of remote, automatic, and passive monitoring for patients with social care needs (known as telecare). The result was the Whole Systems Demonstrator project, funded by the Department of Health, which tested the benefits of integrated care supported by telehealth and telecare in three sites in England (Cornwall, Kent, and Newham).

One of the project's aims was to test the effect of telehealth if delivered at a larger scale than existing pilot schemes, which were often limited to fewer than 100 patients. The resulting trial was pragmatic in design, to recruit and randomise suitably large numbers of patients and assess the effect of a broad class of telehealth and telecare technologies in the context of routine delivery of care provided by the United Kingdom's health service. The telehealth part of the study included people diagnosed with chronic obstructive pulmonary disease, heart failure, or diabetes. These conditions have high prevalence and associated health care costs. The evaluation covered several different dimensions (Bower et al. 2011). This article is one of five analyses, and reports on how telehealth affected the use of secondary health care and mortality. The other analyses will assess how telehealth affected quality of life and cost effectiveness, and explore the patient, professional, and organisation factors related to implementation.

5.4 Methods

The trial protocol has been described by Bower et al. (2011). Initial discussions with sites indicated that individual randomisation of patients would probably not be acceptable to stakeholders. Therefore, we used a pragmatic approach to randomise general practices. Participants in practices

allocated to the control group were given usual care, which reflected the range of services available in the trial sites, excluding telehealth, and were offered telehealth or telecare at the end of the trial, if they were still eligible at that point.

Choices of telehealth devices and monitoring systems varied among the three trial sites, and there was no attempt to standardise these technologies across sites. We included a broad class of technologies, and the study was not designed or powered to examine differences between specific devices or monitoring systems. Although sites used different protocols for allocating peripheral devices, they all used a pulse oximeter for chronic obstructive pulmonary disease, a glucometer for diabetes, and weighing scales for heart failure. Sites asked participants to take clinical readings at the same time each day for up to five days per week, although the frequency was adjusted according to their individual history. For example, a participant with diabetes and well-controlled blood glucose would be asked to take readings less frequently than another participant with poorly controlled blood glucose.

In addition to the telemonitoring aspect of the intervention, symptom questions and educational messages were sent to participants either via the telehealth base unit or via a set top box connected to a television. At the end of each session, data from clinical readings and symptom questions were transmitted to monitoring centres via a secure server. Monitoring centres were staffed by specialist nurses and community matrons from local health organisations, who used protocols to respond to the information from patients.

All practices in the geographic areas covered by three study sites (Cornwall, Kent, and Newham in England) were eligible to participate in the trial. Practices that accepted the invitation to participate were allocated to an intervention or control group via a centrally administered, minimisation algorithm that aimed to ensure that the groups of practices were similar in terms of practice size; deprivation index; proportion of non-white patients; prevalence of diabetes, chronic obstructive pulmonary disease, and heart failure; and site. Within each practice, patients aged 18 years or over were deemed eligible on the basis of a diagnosis in primary or secondary care of chronic obstructive pulmonary disease, diabetes, or heart failure. We did not confer eligibility on the basis of formal clinical assessment of disease severity. Instead, patients were deemed eligible on the basis of their inclusion on the relevant Quality Outcomes Framework register in primary care; a confirmed medical diagnosis in primary or secondary care medical records, as indicated by general practice Read codes or ICD-10 (International Classification of Diseases, 10th revision) codes; or confirmation of disease

status by a local clinician (a general practitioner or community matron) or the patient's hospital consultant. We did not exclude patients on the basis of additional physical comorbidities.

To meet ethical obligations, sites asked patients to complete and return a data-sharing letter if they consented to their data being shared with the research team. Once this letter had been returned, patients received a 'light touch' visit from members of the site project team, sometimes including clinical staff. These visits aimed to assess the suitability of the patient's home for telehealth, provide information regarding the trial, and provide consent forms for participation. The study design required that patients were blinded at the point of consent. However, owing to the large sample size needed for the trial and the extended period of recruitment, it was not possible to guarantee the blinding of recruiters to the allocations of general practices throughout the process.

We assessed the effect of telehealth at the patient level. The primary endpoint was the proportion of people with an inpatient admission to hospital within the 12-month trial period. The study was powered on the basis of detecting a relative change of 17.5% from a baseline of 25% (from *a priori* site estimates), at 80% power, and a two-sided value of $p < 0.05$. Previous studies in the older population suggested that the intracluster correlation coefficient would be about 0.001 (Lancaster et al. 2007). We did sample size calculations using the appropriate formulas (Hayes and Bennett 1999), and found that 3,000 patients would need to be recruited (25 patients from each of 120 general practices). We examined mortality over 12 months and prespecified secondary endpoints (including the number of inpatient bed days, emergency admissions, elective admissions, outpatient attendances, and emergency department visits, as well as the notional cost of hospital activity to commissioners of care based on national tariff costs).

Participants were linked at the person level to data for inpatient and outpatient secondary activity sourced from Hospital Episode Statistics, a national data warehouse for England (The Health & Social Care Information Centre 2013). Participants were linked by the NHS Information Centre for health and social care, a trusted third party that was the only organisation to have access to both patient identifiers and data for secondary care activity. A linked mortality file provided data for all deaths occurring in and out of hospital. In addition, participants were linked to local commissioning datasets on visits to emergency departments, which included all visits to emergency department and not just those that resulted in an admission, and to general practice datasets. In these data, patient identifiable fields were removed before transfer and NHS numbers encrypted. We used this encrypted NHS number to link participants to the emergency and general practice data.

We restricted analysis of inpatient activity to ordinary admissions, and excluded transfers, regular ward attendances, and maternity events (leaving patient classifications 1 and 2 only). Admissions were classified by defined admission methods into emergency activity (codes 21-28) and elective activity (all other codes excluding transfers). Bed days included stays after emergency and elective admissions; same day admissions and discharges were assigned a stay length of 1 bed day. We restricted outpatient activity to appointments that were attended (codes 5 and 6). A separate paper is planned to detail the effect of telehealth on costs.⁴⁹ Here, we included notional costs of hospital care to summarise overall levels of hospital use in the intervention and treatment groups across the inpatient and outpatient categories. We estimated notional costs of care, from Hospital Episode Statistics data, by applying the set of mandatory and indicative tariffs used in England for the reimbursement of inpatient and outpatient care (2008-09 payment by results tariffs) (Department of Health 2007). These tariffs assume a stay of a certain number of days (the 'trim point'), and allow hospitals to charge a prespecified amount for each additional excess bed day. Costs were not adjusted for the regional costs of providing care, and thus were effectively a weighted activity measure that allowed robust comparison of the magnitude of care received for control and intervention participants. We did not include costs for activity not covered by the tariffs, such as mental health, critical care, cystic fibrosis, high cost drugs, and outpatient physiotherapy.

The current study was restricted to those patients linked to administrative data who began the trial before 30 September 2009. The trial start date was taken as the date of telehealth installation for intervention patients, and as the date of the 'light touch' visit for control patients. Analysis was based on comparing activity over 12 months after this date, at the person level.

Analysis of participants was on the basis of the intended treatment allocations, and regardless of subsequent withdrawal from the trial. For randomised trials, formal statistical tests on the similarity of intervention and control patients have been thought to be inappropriate, since allocations are known to have been random (Roberts and Torgerson 1999). However, in cluster randomised trials, selection bias is theoretically possible, either through systematic differences between practices in the control and intervention groups, or because of similar differences at the individual level (Puffer, Torgerson, and Watson 2003). We presented standardised differences as a summary measure of differences between groups; we calculated a standardised difference as the difference between the sample means (or proportions), divided by the pooled standard deviation (Flury and Reidwyl 1986).

⁴⁹ This paper about costs has since been published (Henderson et al. 2013).

Although various aspects of the trial design mitigated against the risk of selection bias, differences between groups could still have occurred by chance. We applied case mix adjustment using three models to account for the effect of any differences between groups (Box 5).

Box 5: Models used for case-mix adjustment

Unadjusted model: The simplest models, although accounting for the effect of clustering, used no additional covariate adjustment.

Adjusted model: A more complex model additionally controlled for residual imbalances in a set of characteristics predictive of future hospital use. These characteristics included age, sex, ethnicity, study site, number of chronic health conditions, principal long-term condition (diabetes, chronic obstructive pulmonary disease, or heart failure), an area-based deprivation score (national quartiles of the Index of Multiple Deprivation 2007), and a metric corresponding to the endpoint calculated over several periods within the two years before recruitment. The number of chronic health conditions was a count of diagnoses recorded on inpatient data over the three years before starting the trial. We assigned principal long-term conditions using a pragmatic approach according to published criteria (Bower et al. 2011).

Combined Model: More complex case-mix adjustment was conducted using the Combined Model (Wennberg et al. 2006). This model is a standard instrument designed to estimate the probability that an individual will be admitted to an emergency hospital department within a 12-month period. The Combined Model score accounts for 72 variables related to age, sex, recorded health conditions, previous hospital use, and prescriptions. These variables are sourced using administrative data from general practices as well as from local hospital commissioning datasets. The Combined Model was originally derived using data for 2002-05 (Wennberg et al. 2006); we revalidated its performance on more recent data covering the period from April 2007 to March 2009. Revalidation used data extracted for the trial sites excluding trial participants. For the case-mix adjustment, we calculated the Combined Model score for each participant at the end of the month before the start date. If a general practice did not grant approval for us to extract data, we imputed scores for patients on the basis of the available information, which included age, sex, and hospital data. We used single imputation on the basis of linear regression on the logit scale.

In a cluster-randomised trial, hospital use for individuals in the same general practice will tend to be correlated. We accounted for this degree of clustering by constructing multilevel models that included random effects at the practice level. Logistic regression was used for the admission proportion and mortality, with the exponent of the coefficients used to calculate odds ratios. For emergency admissions, elective admissions, outpatient attendances, and emergency department visits, we used Poisson regression and exponentiated the coefficients to produce incidence rate ratios. Distributions of health care costs and hospital bed days are typically skewed, with some very large values and a considerable proportion of the population at zero. Although opinions differ on how to analyse such data (Thompson and Barber 2000), we incremented and log transformed notional costs and bed days to meet the assumptions needed for subsequent ordinary least squares modelling. Model coefficients were exponentiated to calculate geometric mean differences. We did all analyses in Stata 11 (StataCorp. 2009).

The primary analysis assumed a 12-month follow-up for all patients, regardless of whether they died or not. This tested for differences between the groups in overall levels of hospital activity after the introduction of telehealth. However, clinicians and other health care professionals might also be interested in how telehealth would affect patients' experiences of admission to hospital, which would depend on whether they were alive at that point. Such effects may be different if telehealth affected the mortality rate. Therefore, we did secondary analyses to assess group differences in admission rates at any point in time, on the condition that participants were alive just before that point in time and had not already been admitted. This analysis treated death as a form of statistical censoring and used the Kaplan-Meier curve (Kaplan and Meier 1958). Although the Kaplan-Meier curve did not take into account differences between intervention and control groups at baseline, we also estimated the corresponding adjusted hazard ratio. We calculated hazard ratios using a Cox proportional hazards model (Cox and Oakes 1984), which included covariate adjustment according to the set of baseline variables (Box 5) and random frailties to allow for homogeneity within practices (Glidden and Vittinghoff 2004).

We calculated full Combined Model scores for 1,397 control patients and 1,365 intervention patients (88% of those included in the analyses), and imputed scores for the remainder of patients. Most inpatient and outpatient hospital activity for the selected patients could be assigned unit costs using our methods. We assigned unit costs to 3,189 (96.3%) of 3,310 inpatient spells experienced by participants during the 12 months before the start of the trial, and to 13,670 (86.7%) of 15,766 outpatient attendances.

5.5.2 Baseline characteristics and trends in hospital activity

Table 13 and Table 14 show that intervention and control patients were similar at baseline (all but one standardised difference <10%). The largest difference between intervention and control patients related to diabetes as an index condition (25.9% *vs.* 21.6%), followed by mean age (69.7 *vs.* 70.9 years). Intervention patients also had less costly hospital activity than controls in the 90 days before the start of the trial (£427 (€529; \$662) *vs.* £506).

Figure 10 shows trends in hospital activity without adjusting for clustering or baseline covariates (that is, as crude data). The break in the chart corresponds to the trial start date, and the chart summarises activity over a series of quarters before and after this date. Rates of emergency hospital admission had peaked for both intervention and control groups at around six quarters before the start of the trial. After the trial began, emergency admissions increased for the control group, from 0.13 per head in the quarter immediately before to 0.18 per head in the quarter immediately after. Emergency department visits also increased over this period, and the rate of conversion of emergency department visits into admissions rose slightly from 53% to 58%. After the initial increase in activity for the control group, rates of emergency admission for the two groups began to converge, although a difference in favour of the intervention group seemed to persist for the entire follow-up period. Table 15 shows summary figures for the crude rates of hospital activity during the follow-up period.

Table 13: Baseline characteristics in the telehealth study

	Control group (n=1,584)	Intervention (n=1,570)	Standardised difference (%)
No of practices	85	92	-
No of patients per practice (median (range))	12 (1-98)	8 (1-77)	-
Index long term condition			
Chronic obstructive pulmonary disease	786 (49.6)	739 (47.1)	-5.1
Diabetes	342 (21.6)	406 (25.9)	10.0
Heart failure	456 (28.8)	425 (27.1)	-3.8
No of chronic health conditions per patient (mean (SD))	1.8 (1.8)	1.8 (1.8)	-3.0
Site			
Cornwall	614 (38.8)	558 (35.5)	-6.7
Kent	576 (36.4)	563 (35.9)	-1.0
Newham	394 (24.9)	449 (28.6)	8.4
Age			
Mean (SD) age in years	70.9 (11.7)	69.7 (11.6)	-9.7
<65 years	446 (28.2)	463 (29.5)	2.9
65-74 years	500 (31.6)	548 (34.9)	7.1
75-84 years	500 (31.6)	446 (28.4)	-6.9
≥85 years	138 (8.7)	113 (7.2)	-5.6
Female	643 (40.6)	647 (41.2)	1.3
Ethnicity			
White	1168 (73.7)	1127 (71.8)	-4.4
Non-white	173 (10.9)	182 (11.6)	2.1
Unknown	243 (15.3)	261 (16.6)	3.5
Area level deprivation ⁵⁰			
Mean (SD)	27.9 (13.5)	28.4 (14.8)	2.9
First quartile	105 (6.6)	137 (8.8)	8.0
Second quartile	272 (17.2)	258 (16.5)	-1.8
Third quartile	571 (36.1)	522 (33.4)	-5.6
Fourth quartile	633 (40.0)	644 (41.3)	2.5
Combined Model score ⁵¹			
Mean (SD)	0.261 (0.200)	0.260 (0.202)	-0.3
Low risk category	226 (16.2)	221 (16.2)	0.0
Moderate risk category	431 (30.9)	440 (32.2)	3.0
High risk category	591 (42.3)	559 (41.0)	-2.7
Very high risk category	149 (10.7)	145 (10.6)	-0.1

Notes: Data are number (%) of patients unless stated otherwise. SD=Standard Deviation.

⁵⁰ n=1,581 (controls), 1,561 (intervention). First quartile is least deprived; fourth quartile is most deprived.

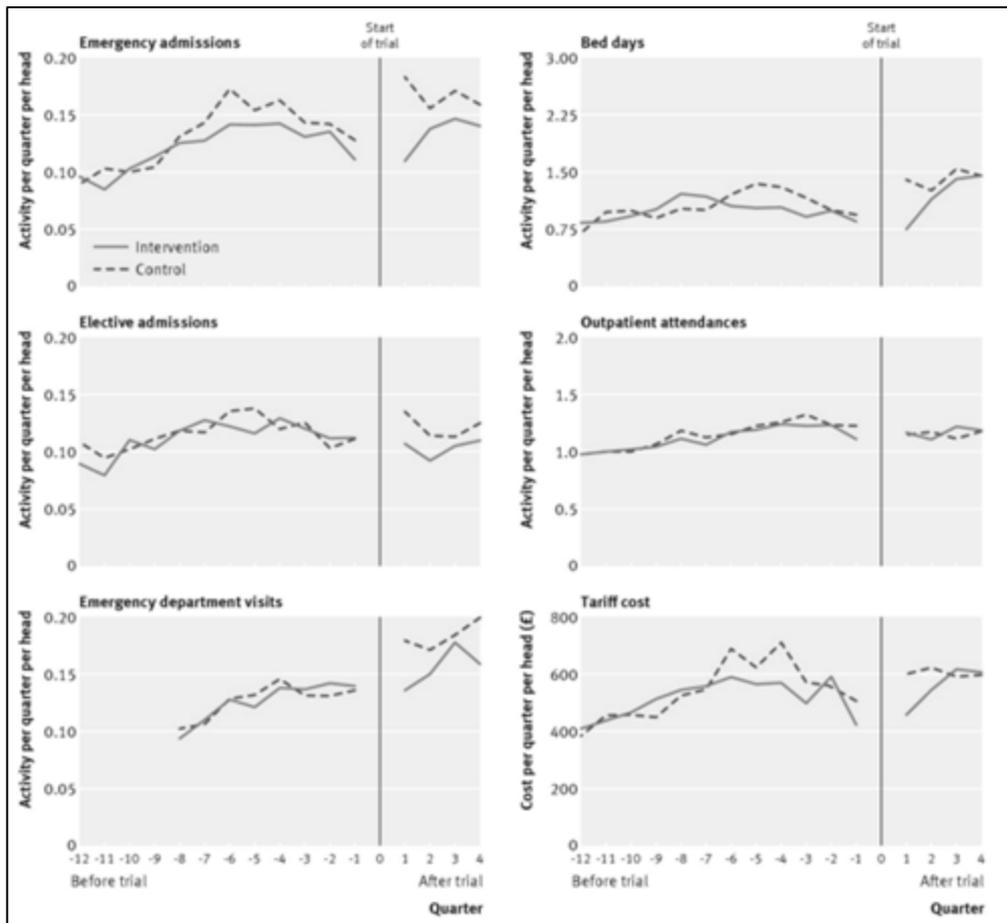
⁵¹ n=1,397 (controls), 1,365 (intervention). Risk categories denote top proportions of site population: very high risk (0.5%), high risk (0.5-5%), moderate risk (5-20%), and low risk (20-100%).

Table 14: Baseline characteristics in the telehealth study (use of secondary care)

Characteristic and period before start of trial	Control group (n=1,584)	Intervention group (n=1,570)	Standardised difference (%)
Admission proportion (%)			
1-90 days	17.5 (n=277)	16.3 (n=256)	-3.2
91-365 days	39.3 (n=622)	39.3 (n=617)	0.1
366-730 days	49.6 (n=785)	47.1 (n=740)	-4.9
Emergency admissions per head			
1-90 days	0.13 (0.45)	0.11 (0.44)	-3.3
91-365 days	0.45 (1.01)	0.42 (0.94)	-3.8
366-730 days	0.60 (1.19)	0.54 (1.07)	-5.5
Bed days per head			
1-90 days	0.94 (4.69)	0.84 (3.81)	-2.3
91-365 days	3.54 (11.25)	3.00 (8.96)	-5.3
366-730 days	4.65 (11.45)	4.57 (12.77)	-0.6
Elective admissions per head			
1-90 days	0.11 (0.40)	0.11 (0.48)	0.5
91-365 days	0.36 (0.95)	0.36 (1.10)	0.6
366-730 days	0.52 (1.29)	0.49 (1.30)	-2.3
Outpatient attendances per head			
1-90 days	1.23 (1.97)	1.14 (1.97)	-4.7
91-365 days	3.87 (5.24)	3.75 (4.98)	-2.4
366-730 days	4.75 (6.35)	4.63 (6.12)	-1.9
Emergency department visits per head			
1-90 days	0.13 (0.44)	0.14 (0.48)	1.1
91-365 days	0.41 (1.05)	0.42 (1.10)	0.9
366-730 days	0.47 (1.05)	0.46 (1.21)	-1.1
Tariff cost per head (£ per head)			
1-90 days	506 (1311)	427 (1177)	-6.3
91-365 days	1879 (3712)	1706 (3116)	-5.1
366-730 days	2411 (4342)	2296 (4143)	-2.7

Note: Data are mean (standard deviation) unless stated otherwise.

Figure 10: Crude trends in secondary care use for patients recruited into the telehealth study



5.5.3 Analysis of primary and secondary endpoints

Of the intervention participants, 42.9% were admitted to hospital during the 12 months of the trial, compared with 48.2% of controls (Table 15). These proportions corresponded to an unadjusted odds ratio of 0.82 (95% confidence interval 0.70 to 0.97, $p=0.017$; Table 16). The odds ratio takes into account clustering at the general practice level. The intraclass correlation coefficient (0.017) was higher than assumed in the original power calculations. The odds ratio for the admission proportion at 12 months remained significant after we adjusted for baseline characteristics and also after we adjusted for the Combined Model score. We assessed the performance of the Combined Model as a discriminatory tool to predict emergency hospital admissions by using data from the trial sites (excluding trial participants). We estimated the area under the receiver operating characteristics curve to be 0.746 ($n=1,523,038$).

Table 15: Rates of secondary care use and mortality during the trial period (unadjusted for clustering and covariates)

	Control group (n=1,584)	Intervention group (n=1,570)	Absolute difference (95% CI)	Percentage difference (95% CI)
Admission proportion (%)	48.2 (n=763)	42.9 (n=674)	-5.2 (-8.7 to -1.8)	-10.8% (-18.1% to -3.7%)
Mortality (%)	8.3 (n=131)	4.6 (n=72)	-3.7 (-5.4 to -2.0)	-44.5% (-65.3% to -23.8%)
Emergency admissions per head	0.68 (1.41)	0.54 (1.16)	-0.14 (-0.23 to -0.05)	-20.6% (-33.8% to -7.4%)
Elective admissions per head	0.49 (1.31)	0.42 (0.99)	-0.07 (-0.15 to 0.01)	-14.3% (-30.6% to 2.0%)
Outpatient attendances per head	4.68 (6.81)	4.76 (6.74)	0.08 (-0.39 to 0.55)	1.7% (-8.3% to 11.8%)
Emergency department visits per head	0.75 (1.58)	0.64 (1.26)	-0.11 (-0.21 to -0.01)	-14.7% (-28.0% to -1.3%)
Bed days per head	5.68 (15.10)	4.87 (14.35)	-0.81 (-1.84 to 0.22)	-14.3% (-32.4% to 3.9%)
Tariff costs per head (£)	2448 (4099)	2260 (4117)	188 (-474.9 to 98.8)	-7.7% (-19.4% to 4.0%)

Note: Data are mean (standard deviation) unless stated otherwise.

Table 16: Estimated treatment effects of telehealth on secondary care use and mortality (results of mixed models, including case-mix adjustment)

Endpoint (interpretation)	Model	Estimate (95% CI)	P-value
Admission proportion (odds ratio)	Unadjusted	0.82 (0.70 to 0.97)	0.017
	Adjusted	0.82 (0.69 to 0.98)	0.026
	Combined Model	0.82 (0.69 to 0.96)	0.016
Mortality (odds ratio) ⁵²	Unadjusted	0.54 (0.39 to 0.75)	<0.001
	Combined Model	0.53 (0.39 to 0.72)	<0.001
Emergency admissions (incidence rate ratio)	Unadjusted	0.81 (0.65 to 1.00)	0.046
	Adjusted	0.85 (0.72 to 1.00)	0.056
	Combined Model	0.81 (0.69 to 0.95)	0.011
Elective admissions (incidence rate ratio)	Unadjusted	0.89 (0.75 to 1.07)	0.219
	Adjusted	0.87 (0.74 to 1.02)	0.078
	Combined Model	0.90 (0.76 to 1.07)	0.241
Outpatient attendances (incidence rate ratio)	Unadjusted	0.96 (0.81 to 1.13)	0.621
	Adjusted	1.01 (0.92 to 1.12)	0.814
	Combined Model	0.95 (0.81 to 1.13)	0.575
Emergency department visits (incidence rate ratio)	Unadjusted	0.85 (0.70 to 1.05)	0.135
	Adjusted	0.85 (0.73 to 1.00)	0.044
	Combined Model	0.86 (0.72 to 1.02)	0.091
Bed days (difference in geometric means)	Unadjusted	-0.64 (-1.14 to -0.10)	0.023
	Adjusted	-0.44 (-0.85 to -0.01)	0.047
	Combined Model	-0.58 (-1.00 to -0.13)	0.013
Tariff costs (difference in geometric means (£))	Unadjusted	-449 (-964 to 243)	0.181
	Adjusted	-242 (-629 to 228)	0.290
	Combined Model	-382 (-840 to 206)	0.184

⁵² One of the models did not converge for mortality.

During the trial, fewer participants died in the intervention group than in the control group (4.6% *vs.* 8.3%; unadjusted odds ratio 0.54, 95% confidence interval 0.39 to 0.75, $p < 0.001$). Differences remained significant after adjustment for the Combined Model score, although we could not adjust for the set of baseline characteristics because models did not converge.

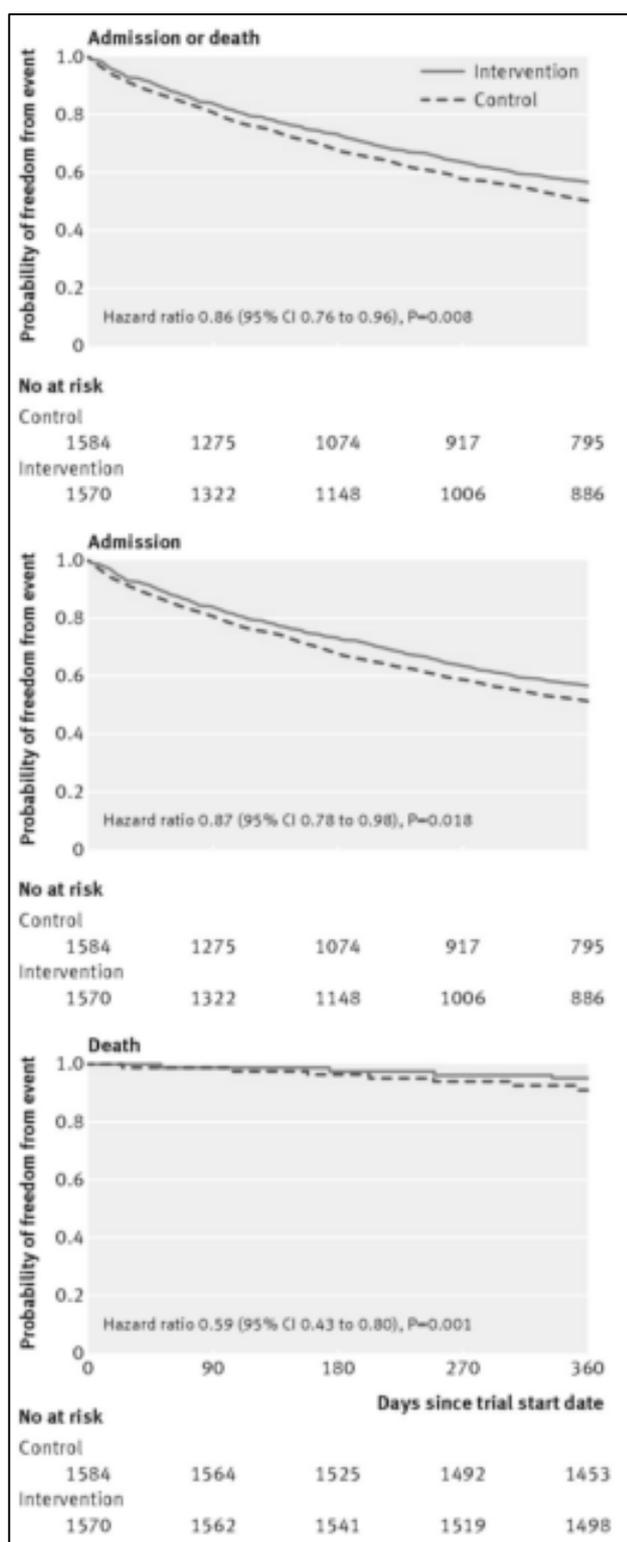
Of the secondary endpoints, emergency admissions, emergency department visits, bed days, and mortality showed significant findings in some or all of the models. Intervention participants underwent 0.54 emergency admissions per head, compared with 0.68 for controls (crude rates), corresponding to an unadjusted incidence rate ratio of 0.81 (95% confidence interval 0.65 to 1.00, $p = 0.046$). However, after we adjusted for baseline characteristics, the upper end of the confidence interval for emergency admissions reached 1.

On average, intervention participants attended emergency departments 0.64 times per head during the trial, compared with 0.75 for controls. This difference was significant in the adjusted estimates only (incidence rate ratio 0.85, 0.73 to 1.00, $p = 0.044$). The intervention and control groups spent an average of 4.87 and 5.68 days in hospital, respectively (unadjusted geometric mean difference -0.64 days, -1.14 to -0.10 , $p = 0.023$); this difference remained significant after adjustment.

Notional costs of hospital activity to commissioners of care were £188 per head lower for intervention participants than for controls (crude rates). Confidence intervals for the geometric mean were very wide and differences were not significant in any of the models (adjusted geometric mean difference $-£242$, 95% confidence interval -629 to 228 , $p = 0.290$).

Secondary analysis of the Kaplan-Meier curves and Cox regression (Figure 11) confirmed that differences in the admission proportion remained significant after censoring observations at death (hazard ratio 0.87, 95% confidence interval 0.78 to 0.98, implying fewer admissions for the telehealth than control group). Graphical methods indicated that the underlying proportional hazards assumption was reasonable. Where Schoenfeld residual tests were significant (Schoenfeld 1982), results remained robust to alternative model specification.

Figure 11: Kaplan-Meier survival analysis for the telehealth study



5.6 Discussion

5.6.1 *Principal findings of the study*

Among a set of patients with chronic obstructive pulmonary disease, diabetes, or heart failure, this study has shown that a smaller proportion of telehealth users than controls were admitted to hospital during a 12-month follow-up. This effect remained significant after adjusting for baseline characteristics and for a predictive risk score. However, the magnitude of the group difference in admission proportion was relatively small (10.8%, 95% confidence interval 3.7% to 18.1%), and smaller than the size that the planned study design was able to detect (17.5%), raising questions about the clinical relevance of the results. The significance of some of the effects reflected the increased power of the study, owing to the higher than assumed baseline level of admissions and to the larger number of small practices (even though the intracluster correlation coefficient was higher than assumed).

Intervention patients were significantly less likely to die within 12 months than controls. We also observed small differences in the mean number of emergency admissions per head between the intervention and control groups (crude rate 0.54 *vs.* 0.68; difference 0.14). These changes were significant in the unadjusted comparisons and when we adjusted for a predictive risk score, but not when we adjusted for baseline characteristics. Hospital bed days were significantly lower among intervention patients than controls, which reflected the reduced admission proportion overall.

For the other measures of hospital use (including the number of elective admissions, outpatient attendances, and emergency department visits), group differences were not significant in general. Crude differences in notional hospital costs to commissioners of care were also not significant and were relatively small (£188 per head over 12 months), especially compared with the potentially high costs of these types of telehealth intervention (Barlow et al. 2007a; Mason et al. 2006; Giordano et al. 2009), which we did not take into account. In view of our results showing confidence intervals crossing the line of no difference, we cannot conclude that telehealth reduces secondary care costs over 12 months. A formal cost effectiveness analysis of the Whole Systems Demonstrator intervention has been undertaken on a subset of participants, using self reported data for hospital use and other services and taking into account the intervention cost, compared with health related quality of life and other outcomes.

Differences in hospital use were at their most marked at the start of the trial, when we observed a distinct increase in admissions for the control group. If we excluded activity from the first three

months of the trial, differences in the admission proportion would not have been significant under any of the models. Therefore, this increase has implications for the interpretation. Trial recruitment processes may have led indirectly to changes in service use for control patients; however, the same processes might also have affected intervention patients in the absence of telehealth. In this case, differences in admissions can be attributed to telehealth, but with the limitation that the trial could have affected the context for the delivery of care for both trial groups.

One explanation for the increase in emergency admissions observed for the control group is that professionals may have identified additional health problems and unmet needs during the recruitment process and could have decided to intervene with control patients not allocated to receive telehealth. This explanation assumes that, if professionals found additional relevant health problems among the intervention group, they were content to manage these in a community setting, with the support of the remote monitoring.

The trial recruitment process might have also raised patients' awareness of their health conditions. Anxiety could have increased as a result of being allocated to the control group, to the extent that these patients were more likely to present at emergency departments and be admitted than intervention patients. The decision to offer telehealth to control patients at the end of the 12-month period, while designed to reduce attrition rates, could have increased anxiety if it encouraged a stronger sense that they were being denied access to support that could be beneficial. A final possible explanation for the increase is that biases could have resulted during patient selection, in which only 35% of patients agreed to the initial 'light touch' visit. There could have been a propensity to select controls with a higher risk of short-term admission and intervention patients with a lower risk. However, observed differences were limited between intervention and control groups, and case mix adjustment was applied.

5.6.2 Strengths and weaknesses of the study

This particular analysis is one of a series planned by the Whole Systems Demonstrator Evaluation Team, and was limited to comparisons of inpatient, outpatient, and emergency department hospital use and mortality. We did not consider the full range of health and social care services, and intervention and control groups could have had differences in the use of primary care, community services, or social care. Telehealth could also have had different effects according to long-term condition or other patient characteristics. Although we assigned unit costs to more than 95% of inpatient activity, we did not consider costs for some elements of hospital care, including mental health and critical care, which had no national tariff. Use of national tariffs meant that the analysis

was relevant to decisions made by commissioners of care, who align with hospital reimbursement guidance, but the economic costs of providing care will differ from the notional costs shown here, and there are regional differences in the costs of providing care.

Although service use could have resource implications, it does not necessarily correlate with health status. Assessment of the effect of interventions should be multidimensional (Fitzpatrick et al. 1992), and important differences could also exist in health outcomes, cost effectiveness, and patient perceptions. These outcomes are explored in the related theme analyses. Telehealth could also have had knock-on effects in non-study patient groups, by freeing up clinical time and resources to care for non-study patients, or by diverting the attention of community teams towards those patients on the trial.

The study used administrative datasets. As a result, person level data were available for 98% of participants. Although these datasets avoided problems of non-response, the quality of data was not directly under the research team's control. Patients tend to underestimate resource use compared with health care providers (Richards, Coast, and Peters 2003), but several studies have pointed out potential problems with using administrative data, such as limited insight into the quality and appropriateness of care (Iezzoni 1997; Roos et al. 1993).

Selection bias is recognised as a risk in cluster randomised trials, in which systematic differences can occur between intervention and control groups at both the cluster and individual level (Puffer, Torgerson, and Watson 2003). At the individual level, if the trial recruiters had foreknowledge of the allocation group (as was often the case here), bias can result through the recruitment of different types of participant into the two groups. We designed this trial to minimise the possibility of bias within the context of a complex community-based intervention. An independent team randomised allocations of practices and a minimisation algorithm aimed to ensure that intervention and control practices were similar in terms of practice size, disease prevalence, and other characteristics. At the individual level, we found no large differences in the characteristics of control and intervention participants at baseline. However, we saw group differences in the median number of participants per general practice (8.5 for telehealth *vs.* 12 for controls). Case-mix adjustment controlled for observed differences between intervention and control groups.

We based this analysis on an intention-to-treat method, which compares patients according to their assigned intervention or control group. Although some patients did not receive their allocated interventions, these numbers were small. A substantial proportion of the intervention group could

have stopped using telehealth before the end of the 12 months. This study had conservative estimates because, in other telehealth applications, equipment might be removed from patients who stop using it (Everett, Kvedar, and Nesbitt 2011). The trial design aimed to minimise differential rates of attrition between intervention and control groups, by ensuring that all practices were allocated to receive a telemonitoring intervention (telehealth or telecare), and that control participants were offered a telemonitoring intervention at the end of the trial period, if they were still eligible.

The effect of telehealth should be considered as just one element within the health system in which it was used. All participating practices and patients in the study could have benefited from the wider service redesign associated with these trials, including those assigned to usual care. Therefore, the study assessed the added value of telehealth over and above the effects of this wider service redesign.

The study aimed to review a broad class of telehealth devices and not to compare specific devices and monitoring systems. Our results reflect specific models of the introduction of telehealth, based on a series of decisions by local teams. There were differences in the interventions offered by the three sites. Although this plurality might be seen as problematic for the purposes of replicating and linking specific aspects of the interventions to likely changes in hospital use, in some ways it is the merit of a pragmatic trial. Other sites introducing telehealth will make choices driven by local contexts in the same way as did the sites in the present study, so the ability to reflect real-life applications of telehealth will add generalisability to the study findings. However, the sites were specifically chosen for their innovations in these areas of care, and conclusions about the effectiveness of telehealth might depend on the environment in which it is used.

5.6.3 Strengths and weaknesses in relation to other studies

Assessment of the effectiveness of telehealth is usually based on assimilating evidence from several small trials, which could make findings difficult to generalise. By contrast, with over 3,100 participants, this study is one of the largest randomised trials of telehealth done so far. The focus on combining three disease groups is novel and allowed us to examine the overall effect of telehealth across populations with chronic disease. However, telehealth could have varying effects in different groups. Other studies have typically examined condition groups separately; thus, comparisons between studies are not straightforward.

Paré and colleagues reviewed 65 empirical studies across four conditions and distinguished the effects of telehealth between different conditions. They suggested that effects on a range of measures (for example, reduced visits to emergency departments, hospital admissions, and average length of

hospital stay) were more consistent in pulmonary and cardiac disorders than in diabetes and hypertension (Paré, Jaana, and Sicotte 2007).

A systematic review of studies for heart failure reported that telemonitoring reduced all-cause mortality, whereas both telemonitoring and structured telephone support reduced admissions for heart failure (Inglis et al. 2010). However, findings were based on generalising a large number of studies with a mean sample size of 330. A study showed that a telehealth intervention that included portable devices significantly reduced costs and admissions for people with heart failure. This was based on a sample size of 460, although the study was adequately powered for the larger effect size it assumed (Giordano et al. 2009). A more recent study of 1,653 patients with heart failure found no significant effect on hospital use or mortality (Chaudhry et al. 2010). In relation to the management of chronic obstructive pulmonary disease, several studies showed lower rates of emergency admissions for patients receiving home monitoring plus telephone support (Bourbeau et al. 2003; Polisen et al. 2010a; Farrero et al. 2001), although one review noted that mortality rates were greater in patients receiving telephone support than those receiving usual care (Polisen et al. 2010a). Evaluations of telehealth interventions for people with diabetes have focused on the achievement of a clinical outcome in terms of glycaemic control (Trief et al. 2009; Welschen et al. 2005),⁵³ with some reported success.

There is also a question about the effect of interventions that combine telemonitoring with educational and motivational tools, such as those we studied, compared with interventions consisting of telemonitoring only. Our study was not designed to answer this question. In a study by Domingo and colleagues of patients with heart failure receiving multidisciplinary care, those who used motivational support tools combined with telemonitoring of weight, blood pressure, and heart rate, thought that their quality of life had improved. These patients also spent less time in hospital after the intervention than before, although before-and-after comparisons can be vulnerable to regression to the mean. Researchers randomised patients to groups with and without the telemonitoring element and found no significant group differences (Domingo et al. 2011).

⁵³ Since the original publication of this paper, there has been an update of the systematic review by Welschen and colleagues (Malanda et al. 2012). The review focused on a more general class of self-monitoring interventions than telehealth, and found small effects on glycaemic control over 12 months. A subsequent analysis of data collected for the Whole Systems Demonstrator evaluation has also found small effects on the same measure for telehealth (Steventon et al. 2014).

5.6.4 Possible explanations and implications for clinicians and policymakers and other researchers

Our results suggest that telehealth helped patients to avoid the need for emergency hospital care. The mechanism for this is not yet clear. Telehealth could help patients manage their conditions better and therefore reduce the incidence of acute exacerbations that need emergency admissions. Telehealth could also change people's perception of when they need to seek additional support, as well as professionals' decisions about whether to refer or admit patients. Further analyses will provide insights into the mechanisms by which telehealth can lead to reductions in admission rates (Bower et al. 2011).

The reduced mortality observed in the intervention group will be an important motivator to invest in these interventions and similar technologies. Although the observed difference in emergency admissions associated with the intervention indicates some potential to reduce use of secondary care, the findings need to be tempered by the estimated scale of the difference in notional hospital cost savings for commissioners of care and the cost of the intervention. Furthermore, the increases seen in emergency admissions among control participants suggest that the trial recruitment processes had an impact. The effect on quality of life must also be considered as part of a broader cost-effectiveness analysis. For commissioners of care services, there are questions about whether any reduction in hospital use for patients receiving telehealth translates to an overall change at the organisational level. Any bed days released as a result could be filled with non-study patients rather than released as cash savings. In turn, this could have meant that health benefits accrued to non-study patients, which were not taken into account here.

The observation of a group effect between intervention and controls could mask differences by subgroups. For local practitioners, it is important to assess whether benefits of telehealth are greater in particular patient types, to inform decisions about prioritising the intervention in specific patient groups. For example, McLean and colleagues observed that telehealth interventions probably did not result in clinically relevant improvements in health outcomes in patients with relatively mild asthma (McLean et al. 2010), but could have a role in patients with more severe disease who are at high risk of admission. The current study was not designed to answer these specific questions.

The effect of telehealth could be intricately linked to wider issues about how health systems operate. It is unclear whether effects are attributable to the technology itself or attributable to how it is implemented (Everett, Kvedar, and Nesbitt 2011), and telehealth could be disruptive because it requires some professional groups to work in different ways.

This analysis is one part of the complete evaluation, and the Whole Systems Demonstrator trial in its entirety will allow a wider discussion of issues around the effects on cost effectiveness, quality of life, and patients' and carers' experiences as well as changes at the organisational level.

Box 6: What this study adds

What is already known on this topic

The known effect of telehealth on secondary health care use for patients with chronic diseases has been based on assimilating findings from several small trials, often with conflicting results.

What this study adds

Among people with chronic obstructive pulmonary disease, heart failure, or diabetes, a broad class of telehealth technologies could be associated with reduced rates of mortality and emergency hospital admission. This effect, however, could be linked to short-term increases in hospital use observed in the control group that may have been affected by recruitment processes during the trial. The estimated scale of hospital cost savings for commissioners of care is modest, and the cost of the telehealth intervention should also be taken into account.

Chapter 6 An empirical assessment of aspects of the generalisability of the Whole Systems Demonstrator trial (research paper 4)

6.1 Preamble to research paper 4

This paper uses placebo tests to examine whether patients in the control arm of the Whole Systems Demonstrator trial experienced similar rates of service use and mortality to a matched subgroup of people who were eligible for the trial but received usual care outside of the trial setting. The paper formulates these placebo tests in terms of specific aspects of generalisability. Thus, a positive result from the placebo tests implies that the control treatment received in the trial was comparable to that received in routine clinical practice and also that any bias resulting from the method used to adjust for differences in patient characteristics between settings was minimal. In this situation, one could consider reweighting the estimate of sample average treatment effect to the characteristics of the target population.

The design of the placebo tests builds on previous work in this area (Hartman et al. 2015; Cole and Stuart 2010). Our approach to variable selection within the placebo tests drew on prior qualitative work into the reasons given by patients for declining to participate in the WSD trial (Sanders et al. 2012), and also previous work that supported the development of the prognostic score used in the WSD study (Chapter 5). The placebo tests used time-series regression (Box 7) as well as generalised linear modelling. The former was identified as a particularly suitable method to estimate effects on emergency admissions, as these events are strongly predicted by previous numbers of emergency admissions (Roland et al. 2005; Billings et al. 2013).

The placebo tests confirm that, in WSD, the RCT control group experienced similar rates of planned forms of health care (*i.e.*, general practice contacts, outpatient attendances and elective admissions) to matched non-participants. However, these tests fail to confirm that the groups experienced similar rates of emergency admission and mortality. Thus, the paper concludes that there is no evidence that the trial is generalisable with respect to these endpoints. A form of sensitivity analysis is developed to illustrate the implications of the failure of the placebo tests. While the main analysis from Chapter 5 reported that telehealth was associated with fewer emergency admissions than usual care, in this

paper I find that the estimated effect of telehealth on emergency admissions differs according to the choice of assumptions about the reasons why the WSD trial is not generalisable.

This paper is forthcoming in *Medical Decision Making* (Steventon, Grieve, and Bardsley, forthcoming). This chapter contains the version of the manuscript that was accepted for publication, after minor formatting and stylistic changes.

Box 7: Time series methods

In time series analysis, the input series is an indicator variable that contains discrete values that flag the occurrence of an event at different points in time. The variable can either correspond to an impulse (if the event has a one-time impact on the response series) or be continuing. A model is specified concerning the relationship between the evolution of the input series and the response. The model coefficients are then fitted to the data to estimate the treatment effect.

One general model for the response series is the autoregressive integrated moving average (ARIMA) model. This contains three elements (termed the autoregressive, integrated and moving average elements) that have non-negative integer orders p , d and q respectively. The model for the response series X_t is as follows, where L is the lag operator and the θ and ϕ terms are parameters of the moving-average and autoregressive elements that must be estimated:

$$\left(1 - \sum_{i=1}^p \phi_i L^i\right) (1 - L)^d X_t = \left(1 + \sum_{i=1}^q \theta_i L^i\right) \varepsilon_t$$

Another approach uses a repeated measures design to predict the response series as a function of time, the input series and covariates. This design specifies the covariance between responses for the same individual at different points in time. In the paper, we used a repeated measures regression and assumed that error terms were autoregressive with order one.

RESEARCH PAPER COVER SHEET

PLEASE NOTE THAT A COVER SHEET MUST BE COMPLETED FOR EACH RESEARCH PAPER INCLUDED IN A THESIS.

SECTION A – Student Details

Student	Adam Steventon
Principal Supervisor	Professor Richard Grieve
Thesis Title	Evaluating complex interventions using routinely collected data: Methods to improve the validity of randomised controlled trials and observational studies

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	n/a		
Have you retained the copyright for the work?*		Was the work subject to academic peer review?	

**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?	Medical Decision Making
Please list the paper's authors in the intended authorship order:	Adam Steventon, Richard Grieve and Martin Bardsley
Stage of publication	In press

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)	I worked with Professor Richard Grieve to design a study that used placebo tests to examine aspects of generalisability more directly, and together we developed the sensitivity analysis that is included in the paper. I structured the data sets, conducted the matching and statistical analyses and produced the first draft of the manuscript, before incorporating comments from the co-authors and peer reviewers.
--	--

Student Signature: _____

Date: _____

Supervisor Signature: _____

Date: _____

6.2 Abstract

Background: Policymakers require estimates of comparative effectiveness that apply to the population of interest, but there has been little research on quantitative approaches to assess and extend the generalisability of Randomised Controlled Trial (RCT) based evaluations. We illustrate an approach using observational data.

Methods: Our example is the Whole Systems Demonstrator (WSD) trial, in which 3,230 adults with chronic conditions were assigned to receive telehealth or usual care. First, we use novel placebo tests to assess whether outcomes were similar between the RCT control group and a matched subset of non-participants who received usual care. We matched on 65 baseline variables obtained from the electronic medical record. Second, we conduct sensitivity analysis to consider if the estimates of treatment effectiveness were robust to alternative assumptions about whether 'usual care' is defined by the RCT control group or non-participants. Thus, we provide alternative estimates of comparative effectiveness by contrasting the outcomes of the RCT telehealth group and matched non-participants.

Results: For some endpoints, such as the number of outpatient attendances, the placebo tests passed, and the effectiveness estimates were robust to the choice of comparison group. However, for other endpoints such as emergency admissions, the placebo tests failed and the estimates of treatment effect differed markedly according to whether telehealth patients were compared with RCT controls or matched non-participants.

Conclusions: The proposed placebo tests indicate those cases when estimates from RCTs do not generalise to routine clinical practice, and motivate complementary estimates of comparative effectiveness that use observational data. Future RCTs are recommended to incorporate these placebo tests and the accompanying sensitivity analyses to enhance their relevance to policy-making.

6.3 Background

Well-conducted Randomised Controlled Trials (RCTs) can ensure high levels of internal validity because the treatment groups are balanced. However, a major concern with RCT evidence is that the resultant estimates may not generalise directly to the target population of interest (Sculpher et al. 2004). Even if an RCT has a pragmatic design without restrictive inclusion criteria and compares the intervention with usual care (Roland and Torgerson 1998), the trial may exclude important subgroups of patients and centres (Gheorghe et al. 2013). Thus both observed and unobserved characteristics that modify treatment effects may differ between the RCT participants and the target population. Another threat to generalisability is that the care provided in the RCT may differ to what would be delivered in routine clinical practice. These concerns about generalisability explain why technologies shown to be beneficial in RCTs may not be diffused into routine practice (Rothwell 2005).

Observational studies, such as prospective cohort studies, have the potential to include a broad range of patients, settings and treatment options, representative of those in routine practice. However, in attempting to estimate comparative effectiveness from observational studies, the major methodological challenge is confounding due to treatment selection (Gross et al. 2006; Cole and Stuart 2010). While there have been recent improvements in methods to deal with confounding (Rosenbaum and Rubin 1983; Sekhon and Grieve 2012), these tend to assume that there is no unobserved confounding, which may be implausible (Kolata 2014). Rather than regarding RCTs and observational studies as mutually exclusive alternatives, a promising research agenda has emerged that uses observational data to assess the generalisability of RCT evidence (Olschewski and Scheurlen 1985; Imai, King, and Stuart 2008; Cole and Stuart 2010; Stuart et al. 2011; Frangakis 2009; Hartman et al. 2015; Pressler and Kaizar 2013).

Cole and Stuart described how treatment effects from an RCT could be generalised by reweighting them to reflect the characteristics of patients in the target population (Cole and Stuart 2010). Stuart and colleagues proposed a diagnostic test that assesses whether patients receiving a given treatment in an RCT reported similar outcomes to those receiving the same treatment in the target population, after reweighting for the characteristics of the two groups (Stuart et al. 2011). Hartman and colleagues formally defined the assumptions required for estimating population treatment effects from RCT data, and proposed accompanying placebo tests (Hartman et al. 2015). None of this previous work has examined how the study should proceed when the placebo tests fail to confirm the generalisability of an RCT. This is an important topic for policymakers because, if a placebo test fails, then it may be unclear whether the intervention is effective in the target population. In this paper, we

extend placebo tests by proposing a sensitivity analysis that addresses the potential implications of non-generalisability. We also apply placebo tests to an RCT of a complex, out-of-hospital intervention (Craig et al. 2008).

Complex interventions are of particular interest to policymakers, and include patient-centred medical homes (Fifield et al. 2013), telehealth (Chaudhry et al. 2010), and health coaching (Wennberg et al. 2010). Placebo tests have not previously been applied to RCTs of these interventions, but such trials face particular threats to generalisability (Craig et al. 2008). For example, the effectiveness of complex interventions is thought to depend on how well teams of care professionals work together, but the presence of an RCT can significantly alter the context in which these teams work (Hendy et al. 2012). Also, patients and professionals often cannot be blinded to treatment allocations in these trials, since the control treatments (*e.g.*, usual care) are typically already known to the participants, and the new interventions require changes in their behaviour (Roland and Torgerson 1998). There are several problems with estimating treatment effects from these non-blinded trials, including 'resentful demoralisation' if a strong preference for the new treatment leads to poor compliance amongst controls (Cook and Campbell 1979; McCambridge, Kypri, and Elbourne 2014).

A clear example of the additional generalisability concerns that are raised by evaluations of complex interventions is the recent Whole Systems Demonstrator (WSD) trial, which used cluster randomisation to assign 3,230 patients with chronic conditions to receive either telehealth or usual care (Bower et al. 2011). In this trial, telehealth was associated with around 20% fewer unplanned ('emergency') hospital admissions than usual care (Steventon et al. 2012a), which led to a national initiative to roll out telehealth and similar approaches to three million people in England (Department of Health 2011). However, the estimated improvements in outcomes following telehealth appeared to be driven by the relative increase in rates of emergency admission among the RCT control group shortly after their recruitment (Steventon et al. 2012a; Hendy et al. 2012), thus raising concern that patients in the control arm reacted to their allocations, or received different care to what would be provided in routine clinical practice.

We describe the WSD trial in the following Section and, in Section 6.5, discuss the generalisability concerns it illustrates. In Section 6.6, we describe placebo tests for assessing whether outcomes differ between the RCT control group and a matched sample from the target population, which we apply to the WSD trial in Section 6.7. In Section 6.8, we propose sensitivity analyses to provide complementary treatment effects to address generalisability concerns. In Section 6.9, we discuss the implications for comparative effectiveness research.

6.4 Running example, the Whole Systems Demonstrator (WSD) trial

The WSD trial used a pragmatic design to assess the impact of telehealth in the context of the routine delivery of care (Department of Health 2006). All primary care practices within three study sites in England (Cornwall, Kent and Newham) were eligible to participate; practices that accepted the invitation (n=369) were randomised according to a minimisation algorithm to provide either telehealth or usual care patients (Bower et al. 2011). Patient inclusion criteria were deliberately broad and specified only age 18 or over plus a diagnosis of Chronic Obstructive Pulmonary Disease (COPD), diabetes or heart failure.

The trial was designed to detect a 17.5% relative change in hospitalisation from a baseline of 25%, at 80% power and a two-sided value of $p < 0.05$ (Bower et al. 2011). The targeted number of patients was 3,000. Potentially eligible patients were identified from the lists of patients registered at the participating primary care practices; diagnoses were sourced from routine primary and secondary care data sets and from clinician reports. Identified patients were written to at home (n=15,171). Those who responded affirmatively (n=5,279) were visited and provided with consent forms for participation. Ultimately, 3,230 patients participated. The treatment allocations of patients followed those of the primary care practices at which they were already registered. While patients could not be blinded, they were only told of their treatment allocations after they had consented to participate. The long recruitment period (May 2008 to September 2009) meant that it was not always possible to blind those recruiting patients.

Telehealth patients received home-based technology to record medical information (for example, blood oxygen) and to answer symptom questions. Information from patients was transmitted automatically to monitoring centres, which were staffed by employees from local health care organisations. Control patients had access to usual care for their area, which did not include telehealth. They were offered telehealth at the end of the 12-month trial period.

For the analysis of service utilisation, primary care practices were asked to share pseudonymised data from the electronic medical record for all their registered patients, covering dates of registrations, encounters, diagnoses, test results and prescriptions over at least a four-year period (Bardsley, Steventon, and Doll 2013). These data were linked to pseudonymised administrative hospital records (Health & Social Care Information Centre 2013).

In pre-specified analyses, telehealth patients experienced fewer emergency hospital admissions than controls over 12 months (incidence rate ratio 0.81, 95% Confidence Interval (CI) 0.65 to 1.00,

$p=0.046$). Differences in other categories of health care utilisation were not statistically significant - this included rates of planned ('elective') admissions, emergency room visits, outpatient attendances, and primary care contacts (Steventon et al. 2012a; Bardsley, Steventon, and Doll 2013). However, intervention patients experienced lower mortality than controls over 12 months. The relative difference, as measured by the odds ratio, was 0.54 (95% CI, 0.39 to 0.75, $p<0.001$), though the absolute change was relatively small (4.6% mortality for telehealth *vs.* 8.3% for usual care).

Although selection bias is often a concern in cluster-randomised trials (Puffer, Torgerson, and Watson 2003), no differences were detected between the baseline characteristics of telehealth and control patients, and effect sizes remained similar after adjustment, suggesting that internal validity was not a major issue. The evaluation protocol pre-specified comparisons between RCT participants and non-participants to consider the generalisability concerns that we now discuss.

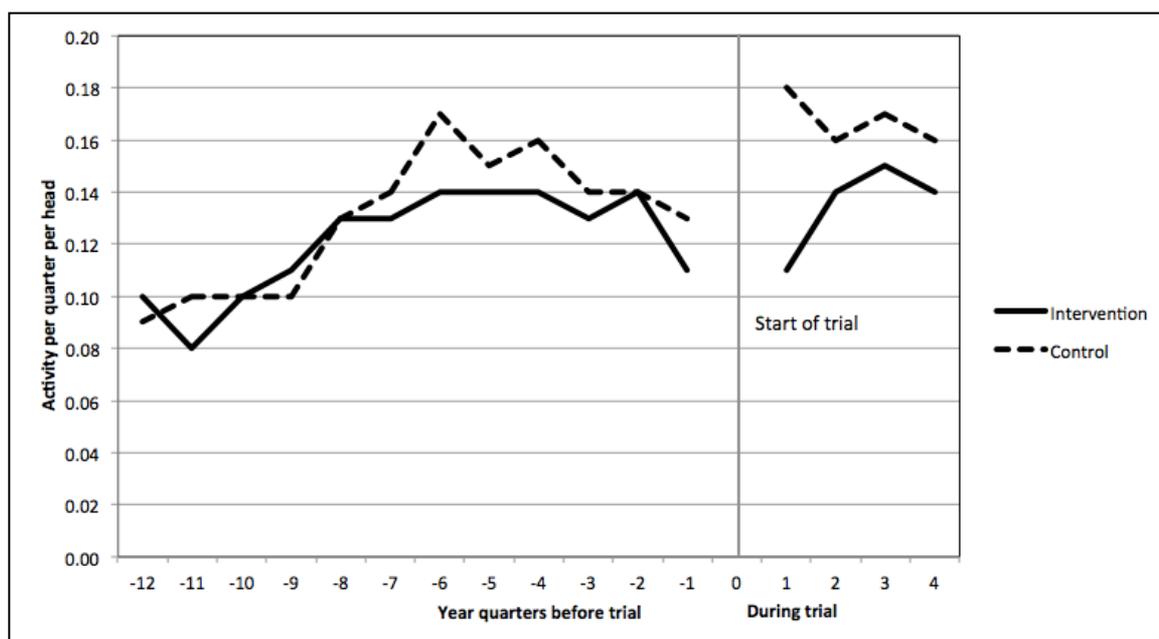
6.5 Concerns about the generalisability of the WSD trial

Concerns about generalisability arose for several reasons. First, as is typical in telehealth trials (Subramanian et al. 2004), only a small proportion of the contacted patients agreed to participate in WSD (20%), suggesting that participants might be unrepresentative of the general population with chronic conditions (Sanders et al. 2012). Second, emergency admission rates increased among control patients shortly after their recruitment (Figure 12) (Steventon et al. 2012a), suggesting that these patients might not have received 'usual care'. Finally, a qualitative study found that the trial protocol and recruitment processes hindered the participating sites' attempts to develop integrated telehealth services (Hendy et al. 2012), suggesting that telehealth might also differ between the RCT and routine practice.

Although the increase in emergency admissions could represent the normal evolution of need for health care among patients with chronic conditions, this seemed unlikely (Hopkinson 2012). Instead, health care professionals might have identified unmet need while recruiting patients and changed the management of those patients allocated to the control group, compared with usual care (Steventon et al. 2012a). Alternatively, the trial recruitment processes might have led to changes in behaviour among patients assigned to the control treatment. Disappointment biases, including resentful demoralisation, have been identified by previous studies of behavioural interventions (McCambridge et al. 2014; Ward et al. 1999). It is possible that some control patients might have felt uneasy or anxious if they perceived that telehealth had benefits that they were excluded from receiving for 12

months, and this contributed to their decision to attend emergency rooms.⁵⁴ Participants can also react strongly to information regarding the likely effectiveness of new treatments, even if this is implied rather than explicit (de Craen et al. 1996). The materials that were provided to patients by the WSD evaluation team were scrutinised by an independent research ethics committee to ensure that they were as neutral as possible, but some of the participating study sites nevertheless advertised and promoted telehealth in local media and through patient advocates (Hobbs 2009).

Figure 12: Patterns of emergency hospital admissions in the telehealth study (n=3,154)



Note: Figure reproduced from the primary WSD study (Steventon et al. 2012a)

Pre-post analysis was not possible for mortality, so there is less direct evidence about the generalisability of this endpoint. However, the mortality effect estimated in WSD (odds ratio 0.54) is larger than other evaluative work in this area (Car, Huckvale, and Hermens 2012). A recent review of systematic reviews concluded that it is probable that mortality from heart failure can be reduced with telemonitoring (McLean et al. 2013), but evidence was less strong for diabetes and COPD, for which meta-analyses have tended to find no effect (McLean et al. 2011; Polisen et al. 2010b, 2009). The effect detected by WSD was across all three conditions, but larger than meta-analyses have found for heart failure alone. It is possible, but unlikely, that an artefact that led to the increases in emergency

⁵⁴ A study that was nested within the WSD trial reported higher anxiety among control than telehealth patients four months after recruitment, but these differences did not reach statistical significance (Cartwright et al. 2013). As this study did not have information on non-participants, it was not possible to test whether this difference was an effect of telehealth or an artifact of the trial.

admissions could also explain the mortality effect. Admissions do have some associated risks, such as adverse events from invasive interventions or hospital-acquired infection (Hauck and Zhao 2011).

Concern about the generalisability of WSD, and thus the benefits of telehealth, appear to explain partly why the spread of telehealth has been limited (Robinson 2013; Takahashi et al. 2012). We will now show how the generalisability of treatment effects can be assessed empirically using placebo tests.

6.6 Statistical methods to assess generalisability

Like most RCTs, WSD reported the treatment effect for patients recruited into the trial, the Sample Average Treatment Effect (SATE) (Imai, King, and Stuart 2008). Equivalent estimands are the Sample Average Treatment effect for the Treated (SATT), which is conditional on assignment to the intervention arm, and likewise for controls (SATC).⁵⁵ However, decision-making usually requires an estimate of the treatment effect for the population that would be eligible for the treatment in routine practice, that is, the Population Average Treatment Effect (PATE), or for those who would receive the treatment in routine practice, that is, the Population Average Treatment effect for the Treated (PATT).

We define RCT results to be 'generalisable' if it is possible to adjust sample average treatment effects to provide an unbiased estimate of the population effect. Hartman and colleagues specified a set of conditions under which such adjustments can be unbiased (Hartman et al. 2015):

- The treatments received in the trial are sufficiently consistent with those in routine practice so as to not have a differential effect on outcomes; and
- There are no unobserved confounders in the selection of the RCT sample from the target population, analogous to the assumption of no unobserved confounding in observational studies (Rubin 2008).

Hartman and colleagues tested these assumptions by comparing outcomes between the RCT and the target population for patients who received the same treatments, using equivalence-based placebo tests (Jones 2007; Hartman and Hidalgo 2011; Hartman, Grieve, and Sekhon 2010). These placebo experiments reverse the hypotheses of standard tests, so that the null hypothesis becomes that the groups have meaningfully different outcomes, while the alternative hypothesis is that the outcomes

⁵⁵ These estimands are asymptotically equivalent because randomisation implies that potential outcomes in the treatment and control groups are exchangeable. They will in general not be equivalent in non-randomised settings.

are similar. Reversing the hypotheses in this way avoids a problem of standard tests, in which lack of evidence for difference can be confused with evidence for similarity. In placebo tests, rejecting the null hypothesis provides support for the generalisability of the trial estimates, as both assumptions A and B will be valid (assuming contrasting effects did not cancel out). In contrast, failure of the placebo tests implies there is no robust evidence to support these assumptions and so, rather than directly reweighting the trial-based estimates (Stuart et al. 2011; Cole and Stuart 2010), alternative approaches may be warranted (see Section 6.8). We next apply placebo tests to the WSD trial.

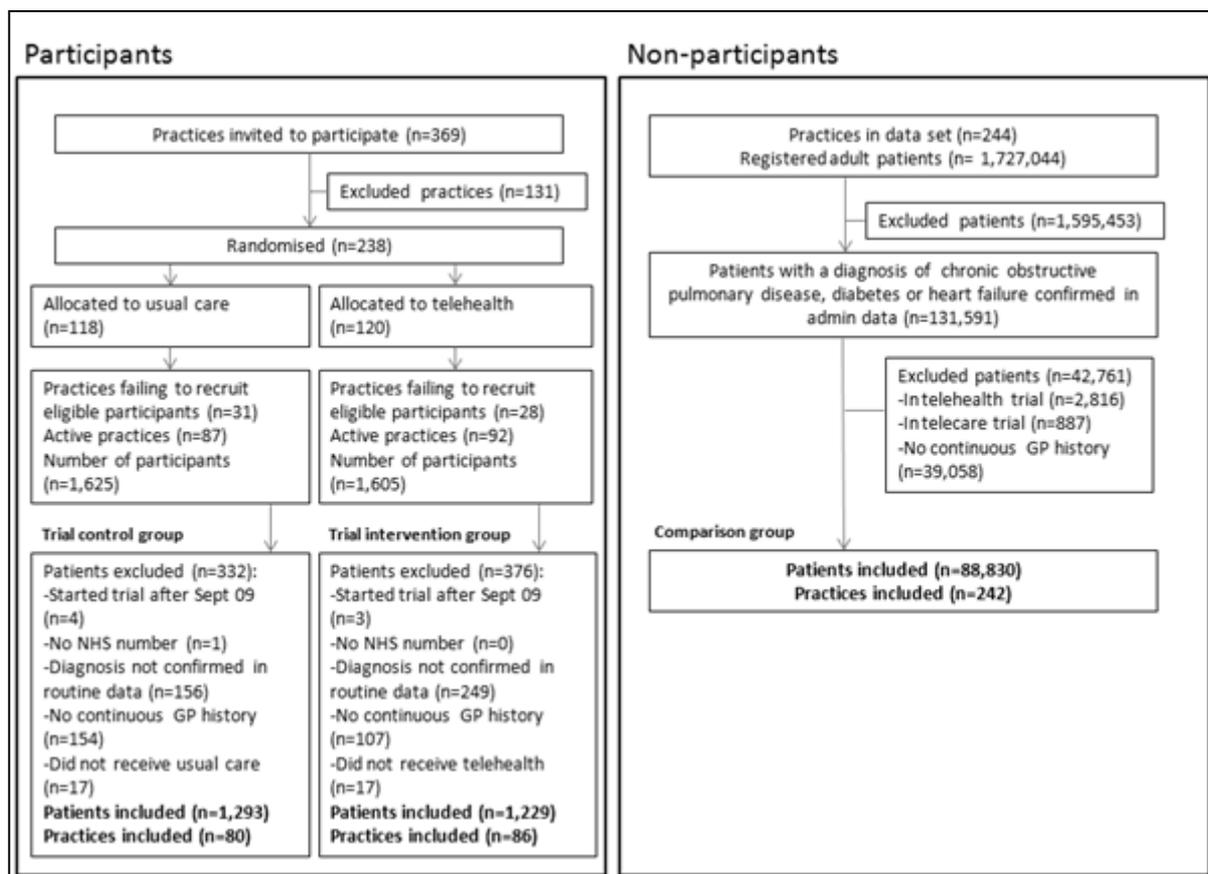
6.7 Applying placebo tests to the WSD trial

6.7.1 Methods

We take the target population to be the wider set of adults in the WSD sites with COPD, diabetes or heart failure, and test whether RCT controls experienced similar outcomes to those predicted by patients in the target population who received usual care. Our method could be readily applied to compare telehealth patients across settings, but we focus on the control group because telehealth has not yet been widely diffused in England. Our definition of 'meaningfully different' is 17.5%, as for the original RCT sample size calculation (Bower et al. 2011).

We identified the target population by applying standard diagnostic codes to the routine primary and secondary care data sets that were collected for the WSD evaluation (n=88,830 eligible non-participants). In order to standardise inclusion criteria across our comparison groups, we excluded the relatively small number (n=405) of RCT participants who had been identified by clinician reports, without a corresponding diagnosis in routine data. Figure 13 shows the flow of patients into the placebo tests. We identified more eligible patients than the original WSD site teams (n=91,647, including participants, compared with 15,171), but our estimate is more consistent with official estimates of disease prevalence (see Appendix B on page 157), and highlights the need to assess the generalisability of the RCT.

Figure 13: Flow diagram showing numbers of general practices and patients available for the placebo tests



Notes: we excluded (from both participant and non-participant groups) people without a continuous record of registration with one or more primary care practices over the two years preceding the trial period. As per the original study, we excluded trial participants who were recruited after September 2009 or not linked to routine data. In the current study, we also excluded the small number of participants who did not receive their allocated treatments.

We identified confounding variables by:

- Drawing on a qualitative study that explored the reasons given by patients for refusal to participate in the WSD trial (Sanders et al. 2012);
- Reviewing existing observational studies of telehealth (see Appendix A on page 157); and
- Considering factors predictive of emergency admissions (Billings et al. 2006).

We identified 65 baseline variables from the routine data, including the Combined Model score, which represented the estimated probability of emergency hospitalisation during the trial period (Wennberg et al. 2006). Other baseline variables related to: demographics; physiological

measurements such as blood pressure; prescription medications; diagnoses of health conditions; prior primary and secondary care use; area-level socioeconomic deprivation; and primary care practice characteristics (see Appendix B for the full list, Table 21, Table 22, and Table 23).

We matched non-participants to RCT control patients using genetic matching, which is a computer-intensive search algorithm that can obtain more closely balanced groups than traditional methods such as pairwise matching on the propensity score (Sekhon and Grieve 2012). While baseline variables were generated for RCT control patients at their enrolment dates, non-participants were assigned multiple 'index dates.' Specifically, each non-participant provided the matching algorithm with up to 14 observations corresponding to month-ends during the trial enrolment period (assuming the non-participant remained eligible for the placebo tests throughout). As any of these observations could be selected as the match for a particular RCT control patient, this approach increased our ability to find well-balanced groups. Genetic matching was applied separately to subgroups of patients defined by study site and chronic condition; within these subgroups, each RCT control patient was matched to one non-participant observation, with replacement. Balance was assessed within and across subgroups using the standardised difference, defined as the difference in sample means as a proportion of the pooled standard deviation (Austin 2008a). A threshold of 10% has been used to describe meaningful imbalances (Normand et al. 2001). We also compared distributions of baseline variables between groups using the variance ratio and quantile-quantile plots (Austin 2009a).

After matched patients had been selected, we calculated utilisation and mortality outcomes over 12 months,⁵⁶ as in the primary studies (Steventon et al. 2012a; Bardsley, Steventon, and Doll 2013). Our main analysis compared each outcome between RCT controls and matched non-participants using generalised linear models, adjusting for residual imbalances in baseline covariates to estimate SATC. The generalised linear models were assumed to follow a Poisson distribution with log link for the utilisation outcomes. Robust standard errors were used to reflect the estimated covariance structure of the data, including the clustering of patients within general practices.

Results from the Poisson regression were presented as incidence rate ratios for RCT controls compared with matched non-participants. The result of each placebo test was reported according to

⁵⁶ One minor difference is that, while the original evaluation used national hospital data from the Hospital Episode Statistics (HES), we used the local equivalent (the Secondary Uses Service, or SUS). SUS enabled us to link anonymised hospital and primary care data across the local populations and not just across WSD participants. Because SUS is the source data for HES, levels of agreement were very high. However, unlike HES, our SUS data would not continue to track hospital activity if a resident of a WSD area moved away from that area.

the confidence interval obtained for the rate ratio. For example, if the point estimate and 95% confidence interval were bounded within a range that corresponded to a 17.5% difference (*i.e.*, within 0.825 to 1.175), the placebo test passed, and we concluded that there was evidence that the RCT control and non-participant groups had similar values of this endpoint. Otherwise, the placebo test failed. For mortality, models used logistic regression rather than Poisson regression and the placebo test was according to the estimated odds ratio.

Matching prior to regression should reduce the sensitivity of the results to the specification of the regression model (Ho et al. 2007), but we also undertook additional robustness checks. These included specifying time-series models to exploit the longitudinal nature of the routine data sets (see Appendix B on page 161).

6.7.2 Results

Before matching, non-participants had less severe case-mix than RCT controls (mean Combined Model score 0.16 *vs.* 0.26, standardised difference 57.2%). After matching, both groups had mean Combined Model score equal to 0.26 (standardised difference 0.3%). Out of the 65 baseline variables, only three had standardised differences that were above the 10% threshold after matching, namely rates of never-smokers, atrial fibrillation and COPD, as recorded in primary care data (standardised differences 11.4%, 10.5% and 10.3%, respectively).⁵⁷ See Table 17 for a summary and Appendix B for detail, page 157). The groups also had similar historic trends in service use (Figure 14).

During the trial, RCT controls experienced similar rates of outpatient attendance, primary care contact and elective admission to matched non-participants (Figure 14). The placebo tests passed for outpatient attendances and primary care contacts – see Table 19, column A. For example, the rate ratio for outpatient attendances was 1.03 (95% CI, 0.94 to 1.13). In contrast, RCT controls experienced more emergency admissions than matched non-participants (rate ratio 1.22, 95% CI, 1.05 to 1.43). They also had higher mortality by 12 months, albeit from a low base (5.7% *vs.* 2.9%; odds ratio 2.17, 95% CI, 1.16 to 4.08). The placebo tests therefore failed for emergency admissions and mortality.

⁵⁷ Atrial fibrillation and COPD showed more similar prevalence on hospital data (standardised differences of 7.4% and 1.4%, respectively), suggesting greater balance was achieved for a subset of patients of more severe case mix than overall. The groups reported similar rates of current smokers and ex-smokers (standardised differences of 2.1% and 8.7%, respectively).

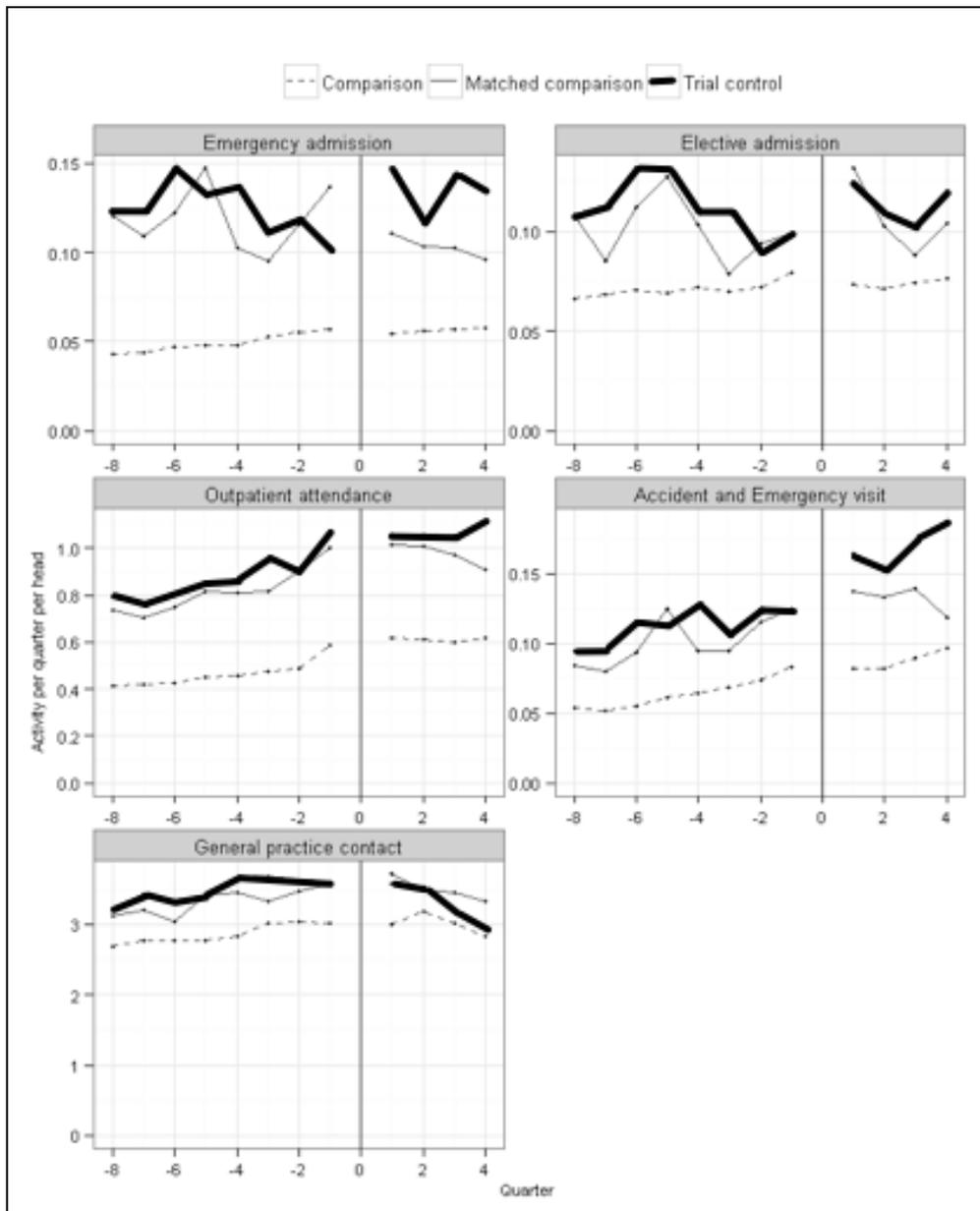
Table 17: Balance, before and after matching, when applying placebo tests to the RCT control group (selected variables)

	Non-participants (n=88,830)	RCT controls (n=1,293)	Matched non-participants (n=1,293)	Standardised difference (variance ratio)	
				Before matching	After matching
Mean practice list size (number of patients per practice ((SD))	9,088 (4,814)	10,041 (5,944)	10,071 (5,758)	17.6 (1.52)	-0.5 (1.07)
Age in years (mean (SD))	66.4 (14.3)	70.8 (11.3)	70.8 (11.1)	34.0 (0.63)	-0.5 (1.04)
Female	46.2	40.3	41.1	-11.9	-1.6
COPD	24.7	60.0	60.2	76.4	-0.3
Diabetes	70.7	34.7	35.5	-77.4	-1.6
Heart failure	12.9	35.2	36.7	54.1	-3.1
Mean number of chronic conditions per head (SD)	0.90 (1.34)	1.76 (1.80)	1.73 (1.77)	54.2 (1.80)	2.1 (1.03)
CM score (mean (SD))	0.16 (0.15)	0.26 (0.20)	0.26 (0.20)	57.2 (1.78)	0.3 (1.00)
Had 10 or more medicines prescribed	7.3	15.9	16.0	27.2	-0.2
Haemoglobin A1c (mean (SD)) ⁵⁸	7.37 (1.63)	8.38 (1.74)	8.23 (1.64)	60.0 (1.13)	8.6 (1.13)
Current smoker	17.6	20.0	19.1	6.0	2.1
Prior number of health care contacts per head, 1-360 days before index date					
Emergency admissions	0.21 (0.67)	0.47 (1.07)	0.45 (0.94)	29.0 (2.60)	2.1 (1.30)
Elective admissions	0.29 (0.99)	0.41 (1.03)	0.38 (0.94)	11.2 (1.07)	3.3 (1.19)
Emergency room visits	0.29 (0.88)	0.48 (1.12)	0.43 (0.94)	19.0 (1.63)	5.1 (1.43)
Outpatient attendances	2.01 (4.01)	3.80 (5.44)	3.53 (4.91)	37.4 (1.84)	5.2 (1.22)
Primary care contacts	11.92 (12.06)	14.55 (12.13)	13.85 (11.00)	21.8 (1.01)	6.1 (1.21)

Notes: data show percentages unless otherwise specified. CM = Combined Model.

⁵⁸ For the diabetes subset only (n=272 intervention patients; 272 matched controls).

Figure 14: Crude trends in health care service use for the placebo tests



Shows mean numbers of contact per patient per quarter. The observations to the left of the vertical line show rates of primary and secondary care contact for each of the eight calendar quarters preceding trial recruitment (*i.e.*, for a total of two years). The observations to the right of the vertical line show rates for survivors for the four quarters in the trial period. A gap has been imposed at the time of recruitment for clarity. The thick black line shows rates for RCT control patients ($n=1,293$). The dashed line shows rates for the eligible non-participants ($n=88,830$), while the thin solid line shows rates for the matched subgroup of eligible non-participants ($n=1,293$). These were matched to RCT controls on variables including prior rates of primary and secondary care contact, and the placebo tests assess whether rates continued to be similar during the trial period. For the purposes of producing this figure, comparison patients were randomly allocated index dates in approximately the same distribution as trial control patients.

Table 18: Rates of health care use and mortality for RCT control and intervention patients and for their corresponding groups of matched non-participants, during the 12 months following index dates

	RCT control group (n=1,293)	Matched non-participants for RCT control group (n=1,293)	RCT intervention group (n=1,229)	Matched non-participants for RCT intervention group (n=1,229)
Emergency admissions per head	0.54 (1.24)	0.41 (0.93)	0.46 (1.02)	0.40 (0.88)
Elective admissions per head	0.46 (1.24)	0.43 (1.09)	0.41 (0.98)	0.50 (2.68)
Outpatient attendances per head	4.28 (6.19)	3.90 (5.20)	4.46 (6.03)	4.05 (5.35)
Emergency room visits per head	0.68 (1.51)	0.53 (1.26)	0.58 (1.21)	0.58 (1.26)
Primary care contacts per head	13.24 (13.08)	14.00 (12.43)	13.57 (12.33)	12.87 (10.79)
Mortality (percentage of patients)	5.7 (n=74)	2.9 (n=37)	2.8 (n=34)	2.1 (n=26)

Note: Data show mean and standard deviation unless otherwise stated.

Table 19: Results of the placebo tests and sensitivity analysis (generalised linear models)

	Placebo tests		Estimated effect of telehealth	
	(A) Comparison of trial control group <i>vs.</i> corresponding matched eligible non-participants	(B) RCT estimate (compares trial intervention group with trial controls)	(C) Sensitivity analysis (compares trial intervention group with corresponding matched eligible non-participants)	
Emergency admissions per head	1.22 (1.05, 1.43)	0.90 (0.77, 1.05)	1.12 (0.95, 1.31)	
Elective admissions per head	0.99 (0.83, 1.18)	0.95 (0.80, 1.14)	0.87 (0.73, 1.05)	
Outpatient attendances per head	1.03 (0.94, 1.13) *	1.02 (0.93, 1.12)	1.04 (0.95, 1.14)	
Emergency room visits per head	1.23 (1.07, 1.43)	0.86 (0.74, 0.99)	0.96 (0.83, 1.11)	
Primary care contacts per head	0.92 (0.87, 0.97) *	1.06 (1.01, 1.13)	1.04 (0.99, 1.09)	
Mortality	2.17 (1.16, 4.08)	0.41 (0.13, 1.23)	1.50 (0.57, 3.94)	

Estimates are for the incidence rate ratio for the trial controls *vs.* comparison populations and accompanying 95% confidence intervals, except for mortality, where odds ratios are reported. * these placebo tests pass, as the point estimate and 95% confidence interval are contained within the range (0.825 to 1.175).

6.7.3 Interpretation of the placebo tests

One explanation for the failure of the placebo tests for emergency admissions and mortality is that the RCT data did not include all confounders of sample selection for these endpoints. Compared with other telehealth studies (see Appendix A on page 157), we were in a relatively strong position to deal with confounding, with 65 baseline variables. We were able to identify retrospectively a group of non-participants that met the inclusion criteria for the WSD trial, and we also had access to a qualitative study of reasons for non-participation in the trial (Sanders et al. 2012). As the electronic medical record was available, we were not confined to non-clinical, administrative data (van Walraven and Austin 2012). Genetic matching produced good balance, both overall and within subgroups defined by study site and chronic condition, thus avoiding confounding through recognising some important interaction effects. Furthermore, our definition of the target population meant that we

selected non-participants from within the same geographical areas as RCT participants. This was expected to reduce confounding due to area-level effects (such as those associated with health care providers) (Steventon, Grieve, and Sekhon 2015), though we also matched on the characteristics of primary care practices. Additional analysis (see Appendix B on page 161) provided reassurance that the placebo tests were not sensitive to the specification of the regression models.

Any remaining confounding variables would need to be unobserved (as good balance was obtained on observed variables) and not strongly predictive of primary care contacts, outpatient attendances and elective admissions (as no differences were seen on those variables). Some candidates are attitudes towards using emergency care and severity of illness requiring emergency care. These, however, are correlated with variables that we did control for, such as prior emergency admissions (Govan et al. 2011), and it is not obvious how they would lead to the sudden increases in admissions that were observed among the RCT control group.

A plausible explanation for the failure of the placebo tests for emergency admissions is that the implementation of the trial protocol altered unplanned forms of care for RCT patients assigned to receive usual care (see Section 6.5). As the placebo tests reported similar rates of elective admission, outpatient attendance and primary care contact (all of which are planned by an appointments system), it now seems unlikely that the health care professionals involved in recruiting patients for the RCT altered the management of patients assigned to the control. However, it is possible that the trial recruitment processes led to changes in behaviour amongst these patients. For example, recruitment into the trial could have increased unease or anxiety amongst those with a preference for telehealth, such that patients were more likely to seek help at emergency rooms (Xu et al. 2008). Unfortunately, information on patient preferences was not collected during the trial (Cook and Campbell 1979), so it is not possible to verify this theory directly.

Compared with emergency admissions, confounding appears more likely for mortality. In-hospital mortality is strongly predicted by clinical information recorded at the point of admission, such as blood pressure and respiratory rate (Quintana et al. 2014; Lee, Austin, and Rouleau 2003), but this information was not available within our data sets at that particular point in the care pathway. It is hard to see how the implementation of the trial protocol could have altered usual care to the extent that mortality rates were significantly increased, though it is true that the mortality effect estimated from the RCT data was unusually large (see Section 6.5).

Regardless of the explanation for the failure of the placebo tests for emergency admissions and mortality, the conclusion is the same: the generalisability of the WSD trial was limited for these endpoints. This encourages sensitivity analyses to determine whether the trial-based estimates are robust to alternative comparison groups.

6.8 Sensitivity analysis regarding reasons for non-generalisability

6.8.1 *Methods*

Failure of the placebo tests implies that either the control treatment in the trial was different to that received in the target population or sample selection was confounded (conditional on observables), or both. Although analysis of RCT data will give unbiased estimates of sample treatment effects, in applying these findings, it is necessary to assume that the treatments are consistent with those in routine practice. We can produce an alternative estimate of the sample treatment effect by comparing the RCT intervention patients with a comparable group of patients receiving the control treatment within the target population. This comparison is observational (and thus susceptible to confounding) but it avoids having to assume that the RCT control group experienced care consistent with usual care in routine practice. We propose that reporting sample treatment effects using both the RCT controls (base case) and a matched sample from the target population (sensitivity analysis), offers a useful check on whether the results from the RCT are robust to these threats to generalisability.⁵⁹

To apply the sensitivity analysis to the WSD study, we matched non-participants to patients in the RCT telehealth arm, using the same approach as before. After satisfactory balance had been obtained (see Appendix C on page 164), we applied generalised linear models to the matched data to estimate the SATT of telehealth versus usual care, by contrasting endpoints for the RCT telehealth arm with the matched non-participant patients who received usual care. We also obtained the comparable estimate from the RCT, by matching controls and telehealth patients within the trial and fitting generalised linear models.

6.8.2 *Results*

For those endpoints where the placebo tests passed, estimated treatment effects were similar regardless of whether telehealth patients were compared with RCT controls (see Table 19, column B), or with matched non-participants (column C). Thus, for example, both analyses found that telehealth did not significantly change the use of outpatient services, and both analyses reported similar point estimates for primary care contacts (though significance levels varied slightly).⁶⁰ In contrast, the estimated treatment effects for emergency admissions and mortality differed markedly according to the comparison group. Thus, although analysis of RCT data alone suggested reductions in emergency

⁵⁹ The proposed sensitivity analysis has some similarities with previous comparisons between the outcomes of people receiving new treatments in RCTs versus those of people receiving control treatments in routine practice (LaLonde 1986). However, these have often been concerned with testing the ability of the analytical methods to adjust for selection bias. We argue that, because the selection mechanism involved relates to recruitment into the RCT, these comparisons are fundamentally related to the generalisability of the RCT.

⁶⁰ The original WSD evaluation found no statistically significant differences in rates of primary care contact (Bardsley, Steventon, and Doll 2013). However, unlike the current study, the original evaluation examined two types of primary care contact separately (those with general practitioners and those with practice nurses). Estimands also differed, and the matching algorithm adopted in the current study may have increased the precision of its estimator.

admissions (rate ratio 0.90, 95% CI 0.77 to 1.05),⁶¹ the comparison between telehealth and matched non-participants reported a trend towards more emergency admissions among telehealth patients (rate ratio 1.12, 95% CI, 0.95 to 1.31). Likewise, analysis of RCT data alone suggested that telehealth reduced mortality (odds ratio 0.41, 95% CI, 0.13 to 1.23), but the comparison with non-participants reported a trend in the opposite direction (odds ratio 1.50, 95% CI, 0.57 to 3.94). These sensitivity analyses have implications for policy-making because reductions in emergency admissions continue to be a major motivation to invest in telehealth.

6.9 Discussion

The current paper adds to a growing body of research addressing the potential of observational data to assess and strengthen the generalisability of RCT-based estimates of comparative effectiveness (Stuart et al. 2011; Cole and Stuart 2010; Hartman et al. 2015). Hartman and colleagues proposed placebo tests for assessing the assumptions required for RCTs to provide unbiased estimates of population average treatment effects (Hartman et al. 2015). We applied placebo tests to an RCT of a complex out-of-hospital intervention, comparing the outcomes of RCT control patients with those of matched non-participants receiving usual care. Unlike previous studies (Stuart et al. 2011; Cole and Stuart 2010; Hartman et al. 2015), we proposed sensitivity analyses to explore the implications of alternative assumptions about the cause of non-generalisability. Sensitivity analyses may be particularly useful for decision-makers in the settings exemplified by our case study, where placebo tests indicate that the main trial findings do not generalise to the target population of interest. For WSD, they showed that, if one assumes that the placebo tests failed because of differences in control treatments between settings, then the data were consistent with an increase in emergency hospital admissions due to telehealth.

Our approach to placebo tests and associated sensitivity analyses has the following requirements. First, before conducting the placebo tests, the target population must be defined, along with what constitutes a meaningful difference in outcomes. In our example, the target population consisted of local patients with the relevant chronic conditions, and the definition of 'clinically different' came from the original sample size calculation.

Second, RCT investigators should understand reasons for non-participation in an RCT. Just as in an observational study, it is vital to understand the mechanism for treatment selection (Rubin 2010). In our example, we drew on a qualitative study about reasons for refusal to participate in the WSD trial (Sanders et al. 2012).

⁶¹ The original analysis of the RCT data reported statistically significant associations between telehealth and emergency admissions, and likewise for mortality. The analysis of RCT data in Table 19 reported non-significant effects for both, but the sample size was smaller.

Third, observational data sets are required with patients, contexts and treatments that represent the target population and overlap with those included in the RCT. These data sets must contain sufficient information to reproduce the main RCT eligibility criteria and outcomes. One possibility is to collect baseline, process and follow-up data for people who refused to participate in the RCT, as in a comprehensive cohort study (Olschewski, Schumacher, and Davis 1992; Brewin and Bradley 1989). Another is to embed RCTs within large routine data sets (van Staa et al. 2012).

Fourth, placebo tests should use analytical techniques that are able to address confounding. We used genetic matching (Sekhon and Grieve 2012), since the large number of baseline variables ($n=65$) was too challenging for traditional matching methods. We also carefully assessed the sensitivity of the placebo tests to alternative regression model specifications (see Appendix B on page 161), including the addition of interaction terms and alternative model structures.

Fifth, as in our example, the study is required to pre-specify sensitivity analysis to assess the robustness of the findings.

Our study has a number of limitations. The placebo tests addressed only some aspects of generalisability, related to the characteristics and treatments of the control group. The generalisability of the intervention regimen could have been addressed with very similar methods, but data were not available on patients receiving telehealth in routine settings. Also, although the placebo tests are useful for providing a quantitative assessment of the generalisability of the RCT-based estimates, they cannot identify precisely why a study was not generalisable. We believe that the most plausible reason for the failure of the placebo tests for emergency admissions in the WSD study was that the control patients received atypical care or otherwise reacted to their treatment allocations, but it is also possible that there were important unmeasured differences in patient characteristics between settings. Finally, as with most comparative effectiveness studies, several endpoints were measured with no attempt to allow for multiplicity. Where the placebo tests indicate that treatment effects were more generalisable for some endpoints than others, judgement is required to decide whether to proceed to estimating population average treatment effects. One explicit approach to weighting the alternative endpoints would be to apply the placebo tests to overall metrics such as those used in cost-effectiveness analyses (for example, to contrast net monetary benefits between the RCT and routine practice settings).

Future research could address alternative estimators for the placebo tests. Stuart and colleagues used three methods to reweight control outcomes to the target population, namely inverse probability of treatment weighting, full matching and sub-classification (Stuart et al. 2011). These gave similar results in their example. We found that generalised linear modelling and time series regression gave similar results (see Appendix B on page 162), but further investigation is warranted. Placebo tests

could also be extended to consider missing endpoints data, and considered for situations without such a rich set of baseline variables as WSD.

This study raises the question about the value of collecting observational data alongside RCTs, given that the amount of funding for a given RCT is constrained and there are competing priorities. Substantial efforts are often made at the design stage to improve the generalisability of RCTs (Roland and Torgerson 1998), but analytical-stage strategies that assess generalisability empirically are often relatively limited. The CONSORT statement, for example, recommends only a table of baseline characteristics (Begg et al. 1966). We conclude that the proposed placebo tests and accompanying sensitivity analyses provide information about the level of generalisability actually achieved in RCTs. These additions to the methodological toolkit can help decision-makers judge the extent to which estimates of comparative effectiveness obtained from RCTs can be adjusted to apply to their target populations.

Table 20: Confounders identified in previous matched control studies of telehealth

	Conditions	Primary outcome	Age	Gender	Ethnicity	Marital status	Living alone	Socio-economics	Principal condition	Secondary conditions	Disease severity	Medication	Hospital use	Primary care use	Patient priority	Predictive risk score
Pekmezaris et al. (2012)	Heart failure	Admissions	Y		Y				Y		Y		Y			
Sohn et al. (2012)	Heart failure	Costs	Y	Y					Y	Y	Y		Y			
Morguet et al. (2008)	Heart failure	Hospital length of stay	Y	Y					Y	Y	Y	Y				
Barnett et al. (2006)	Diabetes	Admissions	Y		Y	Y			Y	Y			Y			
Jia et al. (2009, 2011); Chumbler et al. (2009)*	Diabetes	Various	Y	Y	Y	Y			Y	Y			Y		Y	
Sicotte et al. (2011)	COPD	Various	Y	Y							Y					
Nilsson et al. (2009)	Hypertension	Blood pressure	Y	Y					Y							
Baker et al. (2011)	Various	Health care spending	Y	Y					Y	Y			Y	Y		Y
Chen, Kalish, and Pagan (2011)	Aged over 65	Admissions	Y	Y	Y		Y		Y	Y			Y			
Current study	Mixed	Admissions	Y	Y				Y*	Y	Y	Y**	Y	Y	Y		Y

* At general practice level ** For diabetes

6.10 Appendix A: Confounders analysed in previous matched control studies of telehealth

Table 20 shows the confounders analysed in previous matched control studies of telehealth (remote patient monitoring).

6.11 Appendix B: Matching and placebo tests

6.11.1 Numbers of eligible patients

The WSD site teams identified 15,171 potentially eligible patients using a search of routine primary and secondary care data, whereas we applied standard diagnostic codes to similar data sets and found a much higher number of eligible patients (88,830 non-participants plus 2,817 of participants, *i.e.*, 91,647 in total). Table 22 shows that 24.7% of the non-participants we identified had chronic obstructive pulmonary disease, while 70.7% had diabetes and 12.9% had heart failure.

We sense checked our figures by comparing them with estimates from the system of performance-related pay that exists for general practices in England, the Quality Outcomes Framework (QOF). QOF estimates are based using routine primary care data only and, for diabetes, relate to the population aged 17 or over rather than 18 or over. For 2009/10, QOF reported 35,360, 104,560, and 15,903 people with chronic obstructive pulmonary disease, diabetes, and heart failure, respectively in the three WSD sites. Some people will have several of these conditions.

We would expect the QOF figures to be higher than ours because they are based on all diagnostic codes recorded on the primary care data, however long ago these were recorded. Our primary care data sets, although very extensive with over one billion records, were generally restricted to the period April 2006 to September 2010 (*i.e.*, they started around 3.5 years before recruitment began in September 2009).

Based on the comparison with QOF, we believe that our estimates of the number of eligible patients ($n=91,647$) are plausible. Recruitment into the WSD trial was a complex and time-consuming process that required formal patient consent to be given twice. Various teams had to be carefully coordinated, including the WSD site team, general practices, patients, telehealth installation experts, a market research company (for baseline questionnaires), and social care (for a linked trial of telecare). Therefore, recruitment of patients happened successively for batches of patients, until the target number had been reached ($n=3,000$ across the three sites).

Matching

Baseline variables were calculated as at the trial start date for trial patients and, for non-participants, at up to 14 'index dates' spanning the trial recruitment period (month ends from July 2008 to August

2009). We excluded potential comparison patients who had died before a given index date or whose diagnosis of a chronic condition did not occur before this date.

Matching was performed separately for strata defined by chronic condition and site (*i.e.*, 9 strata in total). Within each stratum, one matched comparison patient was selected for each trial patient with replacement, using Genetic Matching (Sekhon and Grieve 2012). This is a computer-intensive approach that searches over a space of distance measures. In this context, a distance measure evaluates the similarity of two patients at baseline, as a function of their baseline variables. Various measures are possible, with different weights attached to the constituent variables. For each sampled measure, the genetic matching algorithm assembles matched pairs and assesses balance on baseline variables. Thus, the algorithm attempts to find the distance measure that gives the optimal level of balance. Balance was assessed by p-values from paired t-tests as well as, for continuous variables, p-values from Kolmogorov-Smirnov tests. The genetic matching algorithm was run with a population size of 2,000, and stopped after no improvement was detected within 100 generations.

Matched comparison patients selected across the strata were then recombined for the analyses. Means and variances are reported for practice level variables in Table 21. Table 22 and Table 23, show the equivalent figures for the person-level variables, while quantile-quantile plots of the three continuous variables are shown in Figure 15. Most baseline variables had no missing data. For baseline variables for which there was some missing data (such as blood pressure readings), balance statistics for that variable are reported for complete cases.

Table 21: Balance, before and after matching, when applying placebo tests to the RCT control group (practice-level variables)

	Non-participants (n=88,830)	Trial controls (n=1,293)	Matched non-participants (n=1,293)	Standardised difference (variance ratio)	
				Before matching	After matching
Mean practice list size (number of patients per practice ((SD))	9,088 (4,814)	10,041 (5,944)	10,071 (5,758)	17.6 (1.52)	-0.5 (1.07)
Diabetes prevalence	5.7	5.6	5.7	-2.8	-5.5
COPD prevalence	1.6	1.6	1.6	-1.6	-0.5
Heart failure prevalence	0.8	0.8	0.8	12.7	-3.0
Mean socioeconomic deprivation score (SD)	23.8 (9.9)	26.9 (9.7)	26.5 (10.4)	32.0 (0.96)	3.9 (0.87)

Note: Weighted by the sample size for each practice.

Table 22: Balance, before and after matching, when applying placebo tests to the RCT control group (person-level variables)

	Non-participants (n=88,830)	Trial controls (n=1,293)	Matched non-participants (n=1,293)	Standardised difference (variance ratio)	
				Before matching	After matching
Mean age in years (SD)	66.4 (14.3)	70.8 (11.3)	70.8 (11.1)	34.0 (0.63)	-0.5 (1.04)
Female	46.2	40.3	41.1	-11.9	-1.6
COPD	24.7	60.0	60.2	76.4	-0.3
Diabetes	70.7	34.7	35.5	-77.4	-1.6
Heart failure	12.9	35.2	36.7	54.1	-3.1
Mean CM score (SD)	0.16 (0.15)	0.26 (0.20)	0.26 (0.20)	57.2 (1.78)	0.3 (1.00)
Number of distinct medicines					
1 to 4	41.8	28.0	27.9	-29.3	0.2
5 to 9	42.7	45.2	46.2	5.0	-2.2
10 +	7.3	15.9	16.0	27.2	-0.2
Haemoglobin A1c (mean (SD)) ⁶²	7.37 (1.63)	8.38 (1.74)	8.23 (1.64)	60.0 (1.13)	8.6 (1.13)
Systolic blood pressure ⁶³	133.67 (16.46)	132.38 (17.57)	133.30 (17.70)	-7.6 (1.14)	-5.2 (0.99)
Diastolic blood pressure ⁶⁴	75.94 (9.67)	75.14 (10.16)	75.01 (10.62)	-8.1 (1.10)	1.2 (0.92)
Body mass index ⁶⁵	29.66 (6.27)	28.14 (6.08)	28.69 (6.02)	-24.6 (0.94)	-9.1 (1.02)
Smoking status					
Current smoker	17.6	20.0	19.1	6.0	2.1
Ex-smoker	37.9	53.8	49.4	32.3	8.7
Never smoked	44.5	26.2	31.4	-38.9	-11.4
Prior numbers of health care contacts per head over various periods of time before index date					
1-360 days					
Emergency admissions	0.21 (0.67)	0.47 (1.07)	0.45 (0.94)	29.0 (2.60)	2.1 (1.30)
Elective admissions	0.29 (0.99)	0.41 (1.03)	0.38 (0.94)	11.2 (1.07)	3.3 (1.19)
Emergency room visits	0.29 (0.88)	0.48 (1.12)	0.43 (0.94)	19.0 (1.63)	5.1 (1.43)
Outpatient attendances	2.01 (4.01)	3.80 (5.44)	3.53 (4.91)	37.4 (1.84)	5.2 (1.22)
Primary care contacts	11.92 (12.06)	14.55 (12.13)	13.85 (11.00)	21.8 (1.01)	6.1 (1.21)
361-720 days					
Emergency admissions	0.18 (0.62)	0.53 (1.09)	0.50 (1.02)	38.9 (3.06)	2.6 (1.13)
Elective admissions	0.27 (0.96)	0.48 (1.26)	0.43 (1.03)	18.8 (1.74)	4.5 (1.48)
Emergency room visits	0.22 (0.78)	0.42 (0.96)	0.38 (0.97)	22.2 (1.52)	3.5 (0.98)
Outpatient attendances	1.71 (3.61)	3.21 (4.70)	3.00 (4.14)	35.7 (1.69)	4.6 (1.29)
Primary care contacts	11.00 (11.39)	13.35 (11.19)	12.76 (10.48)	20.9 (0.96)	5.5 (1.14)

Notes: Data show percentages unless otherwise specified. CM = Combined Model.

⁶² For the diabetes trial subset only (n=272 intervention patients and 272 matched controls).

⁶³ n=86,519, 1,263, and 1,254, for the three groups, respectively.

⁶⁴ n=86,512, 1,263, and 1,254, for the three groups, respectively.

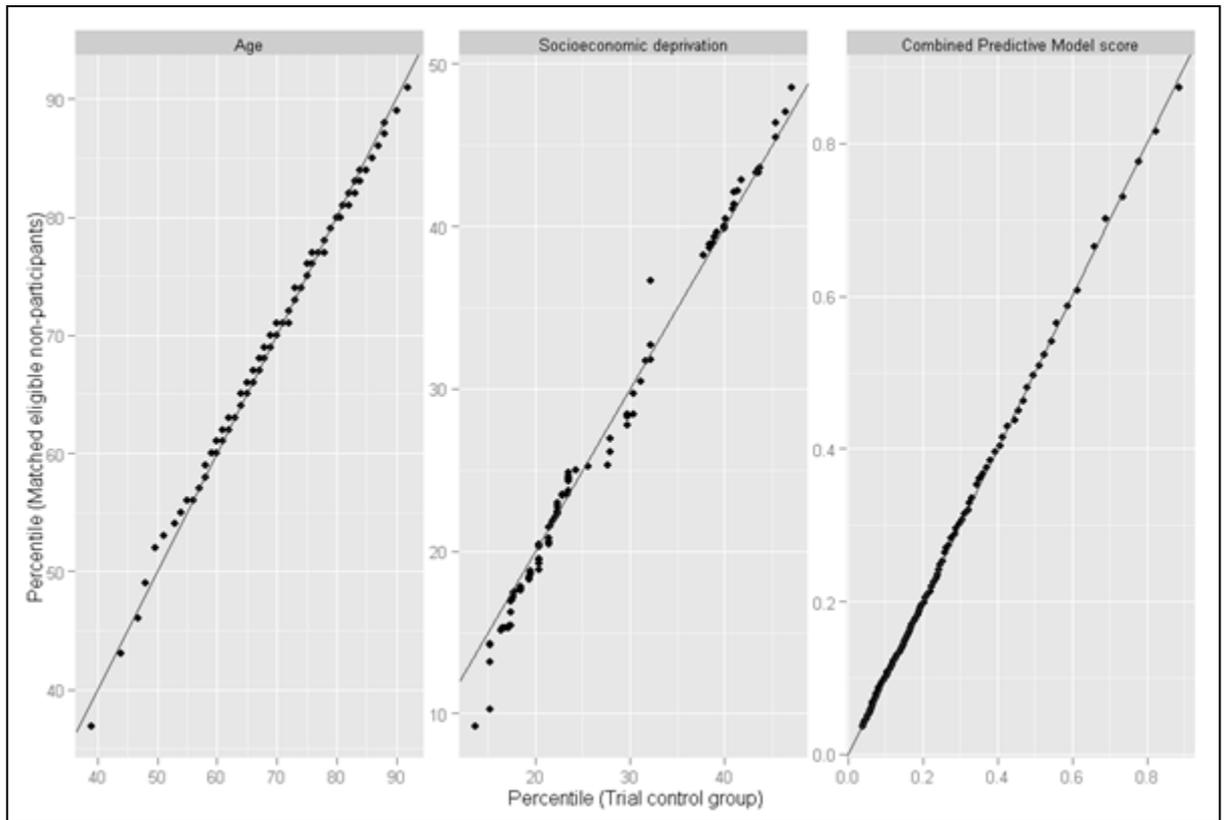
⁶⁵ n=65,552, 1,007, and 882, for the three groups, respectively.

Table 23: Balance, before and after matching, when applying placebo tests to the RCT control group (person-level variables, continued)

	Non-participants (n=88,830)	Trial controls (n=1,293)	Matched non-participants (n=1,293)	Standardised difference (variance ratio)	
				Before matching	After matching
Health conditions recorded on hospital data					
COPD	8.0	26.9	26.3	51.4	1.4
Congestive heart failure	5.7	16.9	17.5	36.0	-1.4
Diabetes	24.2	24.7	24.7	1.4	0.2
Cancer	4.9	8.7	7.7	15.3	3.9
Cancer (benign)	1.4	2.2	2.2	5.7	-0.5
Alcohol abuse	0.8	2.2	1.7	11.8	3.9
Hypertension	22.7	37.5	39.5	32.7	-4.1
Injury	8.3	13.2	14.3	15.8	-3.1
Iatrogenic	2.9	4.9	6.3	10.6	-6.0
Falls	3.4	4.6	3.9	6.3	3.4
Mental health	3.6	5.9	6.6	10.8	-2.9
Angina	5.7	13.8	13.3	27.6	1.6
Ischemic heart disease	10.9	28.4	24.1	45.3	9.9
Asthma	5.8	12.3	11.5	22.9	2.4
Anaemia	3.5	6.8	6.6	15.2	0.9
Atrial fibrillation	6.9	17.2	14.5	32.1	7.4
Cerebrovascular disease	3.3	4.9	6.9	8.2	-8.2
Peripheral vascular disease	3.1	7.3	8.0	19.4	-2.3
Renal failure	2.7	7.3	6.2	21.1	4.3
Respiratory infection	2.8	8.0	5.8	22.8	8.6
Number of chronic conditions per head (mean (SD))	0.90 (1.34)	1.76 (1.80)	1.73 (1.77)	54.2 (1.80)	2.1 (1.03)
Health conditions recorded on primary care data					
Atrial fibrillation	8.4	18.3	14.5	29.4	10.5
Cancer	7.2	10.9	10.4	13.1	1.5
Coronary heart disease	19.7	35.0	33.4	35.0	3.4
Chronic kidney disease	17.6	22.5	23.7	12.4	-2.8
COPD	22.0	56.5	51.4	75.6	10.3
Dementia	1.5	0.8	1.6	-6.5	-7.8
Depression	14.8	17.8	17.0	8.1	2.0
Diabetes	67.8	33.5	33.8	-73.1	-0.7
Heart failure	10.0	31.6	28.8	55.4	6.1
Hypertension	52.7	52.5	52.8	-0.4	-0.6
Mental health	1.6	0.8	1.8	-7.5	-9.0
Asthma	13.8	21.3	22.8	19.7	-3.7
Stroke / TIA	7.1	11.8	11.3	16.0	1.5

Notes: Data show percentages unless otherwise specified. TIA = Transient Ischaemic Attack.

Figure 15: Quantile-quantile plots showing balance when applying placebo tests to the RCT control group



6.11.2 Generalised linear regression

After matched data sets had been assembled, analytical models were applied. Regression models were constructed using generalised linear modelling with a log link, and robust standard errors. The headline results presented in the paper are adjusted for a set of 51 baseline variables; however, we tested a range of other model specifications to assess the robustness of the results. The parsimonious model adjusted for 25 variables selected manually by removing and adding variables in such a way as to optimise model fit according to the Akaike Information Criterion (AIC). More extensive models included interaction terms with age, site and indexed chronic condition, again selected with reference to the AIC. Results for the estimated difference in emergency hospital admissions were robust to changes in model specification on the matched data sets, with only small changes detected in point estimates and confidence intervals (Table 24).

Table 24: Sensitivity of the placebo test to alternative specifications of the generalised linear models (emergency hospital admissions)

Model	Number of adjustment variables, excluding group	Incidence rate ratio (95% confidence interval)
Full model	51	1.22 (1.05 to 1.43)
Parsimonious model	25	1.24 (1.06 to 1.44)
Including interaction terms with age	32	1.22 (1.05 to 1.43)
With site and interaction terms with site	41	1.25 (1.07 to 1.45)
With interaction terms with condition	53	1.27 (1.09 to 1.47)
Predictive risk score	1	1.31 (1.11 to 1.54)

Note: All models were applied to the matched data set.

6.11.3 Time series analysis

In addition to the generalised linear models, we also considered addressing residual confounding after matching by applying repeated measures (“time series”) models to the matched data. These time series models used quarterly utilisation totals spanning the period from two years before the date of enrolment into the trial (or the index date, for non-participants) to 12 months afterwards, and included an interaction term between group membership and trial period. Error terms were assumed to be autoregressive with order one (Cowpertwait and Metcalfe 2009); robust standard errors were applied. Compared with the generalised linear models, the time series models were expected to reduce the risk of unobserved confounding, provided that the unobserved confounder did not have an effect that varied with time. However, it was not possible to apply time series models to the mortality data, as no deaths occurred in the study cohorts before the index dates. Results from the time series models were similar to those from generalised linear modelling (Table 25). Effect sizes from the time series models were also robust to alternative model specification (Table 26).

Table 25: Results of the placebo tests (comparison between generalised linear models and time series analysis)

	Generalised linear modelling	Time series
Emergency admissions per head	1.22 (1.05, 1.43)	1.31 (1.11, 1.54)
Elective admissions per head	0.99 (0.83, 1.18)	0.97 (0.79, 1.19)
Outpatient attendances per head	1.03 (0.94, 1.13)	1.02 (0.93, 1.12)
Emergency room visits per head	1.23 (1.07, 1.43)	1.18 (1.00, 1.41)
Primary care contacts per head	0.92 (0.87, 0.97)	0.92 (0.87, 0.98)
Mortality	2.17 (1.16, 4.08)	n/a

Notes: Figure are for the incidence rate ratio for the trial controls vs. non-participants and 95% confidence intervals, except for mortality, where figures are for the odds ratio. Time series models were not defined for mortality, as no deaths occurred before the date of enrolment into the trial.

Table 26: Sensitivity of the placebo test to alternative specifications of the time series models (emergency hospital admissions)

Data level	Fixed effects	Covariance structure	Incidence rate ratio (95% confidence interval)
Quarterly	Quarterly dummies, group membership and interaction term	First order autoregressive	1.31 (1.11 to 1.54)
Quarterly	Quarter expressed as continuous variable, group membership and interaction term	First order autoregressive	1.25 (1.08 to 1.43)
Quarterly	Quarterly dummies, group membership and interaction term	Spatial	1.31 (1.11 to 1.54)
Monthly	Monthly dummies, group membership and interaction term	First order autoregressive	1.25 (1.06 to 1.48)

6.12 Appendix C: Analysis for trial intervention group

Table 27, Table 28 and Table 29 show the results of the matching algorithm when applied to the trial intervention group. Figure 16 shows trends in health care use for the RCT intervention group and their corresponding matched eligible non-participants. Table 30 shows estimated treatment effects using both generalised linear modelling and the alternative modelling approach based on time series analysis (see Appendix B for a description of the time series modelling).

Before matching, non-participants had less severe case-mix than RCT intervention patients (mean Combined Model score 0.16 *vs.* 0.26, standardised difference 57.4%). After matching, both groups had mean Combined Model score equal to 0.26 (standardised difference 0.9%). Out of the 65 baseline variables, six had standardised differences above the 10% threshold after matching, namely diastolic blood pressure and rates of current smoking, ex-smoking, atrial fibrillation, dementia, and mental health conditions recorded on primary care data (10.5%, -12.1%, 18.0%, 11.9%, -14.4% and -16.4%). The groups also had similar historic trends in service use (Figure 16).

Table 27: Balance, before and after matching, in the sensitivity analysis as applied to the RCT intervention group (practice-level variables)

	Non-participants (n=88,830)	Trial intervention patients (n=1,229)	Matched non-participants (n=1,293)	Standardised difference (variance ratio)	
				Before matching	After matching
Mean practice list size (number of patients per practice ((SD))	9,088 (4,814)	9,417 (3,828)	9,265 (3,719)	7.6 (0.63)	4.0 (1.06)
Diabetes prevalence	5.7	5.8	6.0	13.1	-8.9
COPD prevalence	1.6	1.5	1.5	-10.4	-3.2
Heart failure prevalence	0.8	0.8	0.8	21.1	4.8
Mean socioeconomic deprivation score (SD)	23.8 (9.9)	29.2 (11.5)	28.7 (11.2)	50.4 (1.35)	4.5 (1.05)

Note: Weighted by the sample size for each practice.

Table 28: Balance, before and after matching, in the sensitivity analysis as applied to the RCT intervention group (person-level variables)

	Non-participants (n=88,830)	Trial intervention patients (n=1,229)	Matched non- participants (n=1,293)	Standardised difference (variance ratio)	
				Before matching	After matching
Mean age in years (SD)	66.4 (14.3)	69.1 (11.6)	69.5 (11.3)	21.2 (0.65)	-3.3 (1.06)
Female	46.2	41.0	41.4	-10.4	-0.8
COPD	24.7	54.7	54.2	64.3	1.0
Diabetes	70.7	44.8	45.3	-54.5	-1.1
Heart failure	12.9	33.5	33.8	50.4	-0.7
Mean CM score (SD)	0.16 (0.15)	0.26 (0.20)	0.26 (0.20)	57.4 (1.83)	0.9 (1.01)
Number of distinct medicines					
1 to 4	41.8	28.6	28.4	-28.0	0.4
5 to 9	42.7	47.0	47.2	8.7	-0.3
10 +	7.3	20.3	20.3	38.2	0.0
Haemoglobin A1c (mean (SD)) ⁶⁶	7.37 (1.63)	8.55 (1.95)	8.44 (1.92)	65.6 (1.42)	5.5 (1.03)
Systolic blood pressure ⁶⁷	133.67 (16.46)	130.99 (17.05)	132.35 (17.56)	-16.0 (1.07)	-7.9 (0.94)
Diastolic blood pressure ⁶⁸	75.94 (9.67)	74.01 (10.03)	75.05 (9.85)	-19.6 (1.07)	-10.5 (1.03)
Body mass index ⁶⁹	29.66 (6.27)	29.43 (6.66)	29.08 (6.36)	-3.6 (1.13)	5.3 (1.10)
Smoking status					
Current smoker	17.6	15.9	20.5	-4.7	-12.1
Ex-smoker	37.9	52.6	43.7	29.9	18.0
Never smoked	44.5	31.5	35.8	-27.0	-9.1
Prior numbers of health care contacts per head over various periods of time before index date					
1-360 days					
Emergency admissions	0.21 (0.67)	0.42 (0.88)	0.41 (0.83)	26.7 (1.73)	1.2 (1.12)
Elective admissions	0.29 (0.99)	0.44 (1.25)	0.41 (1.20)	12.5 (1.58)	1.8 (1.08)
Emergency room visits	0.29 (0.88)	0.55 (1.18)	0.51 (1.10)	25.4 (1.79)	4.1 (1.13)
Outpatient attendances	2.01 (4.01)	3.92 (5.09)	3.55 (4.33)	41.7 (1.61)	7.9 (1.38)
Primary care contacts	11.92 (12.06)	13.14 (10.72)	12.93 (10.39)	10.7 (0.79)	2.0 (1.06)
361-720 days					
Emergency admissions	0.18 (0.62)	0.48 (0.96)	0.43 (0.93)	37.5 (2.38)	5.6 (1.07)
Elective admissions	0.27 (0.96)	0.46 (1.24)	0.40 (1.16)	16.9 (1.69)	5.5 (1.16)
Emergency room visits	0.22 (0.78)	0.45 (1.10)	0.42 (1.05)	24.1 (1.98)	2.6 (1.08)
Outpatient attendances	1.71 (3.61)	3.33 (4.38)	3.15 (4.18)	40.4 (1.47)	4.3 (1.10)
Primary care contacts	11.00 (11.39)	12.65 (11.01)	12.18 (10.28)	14.8 (0.93)	4.4 (1.15)

Notes: Data show percentages unless otherwise specified. CM = Combined Model.

⁶⁶ For the diabetes trial subset only (n=346 intervention patients and 346 matched controls).

⁶⁷ n=86,519, 1,204, and 1,205, for the three groups, respectively.

⁶⁸ n=86,512, 1,204, and 1,205, for the three groups, respectively.

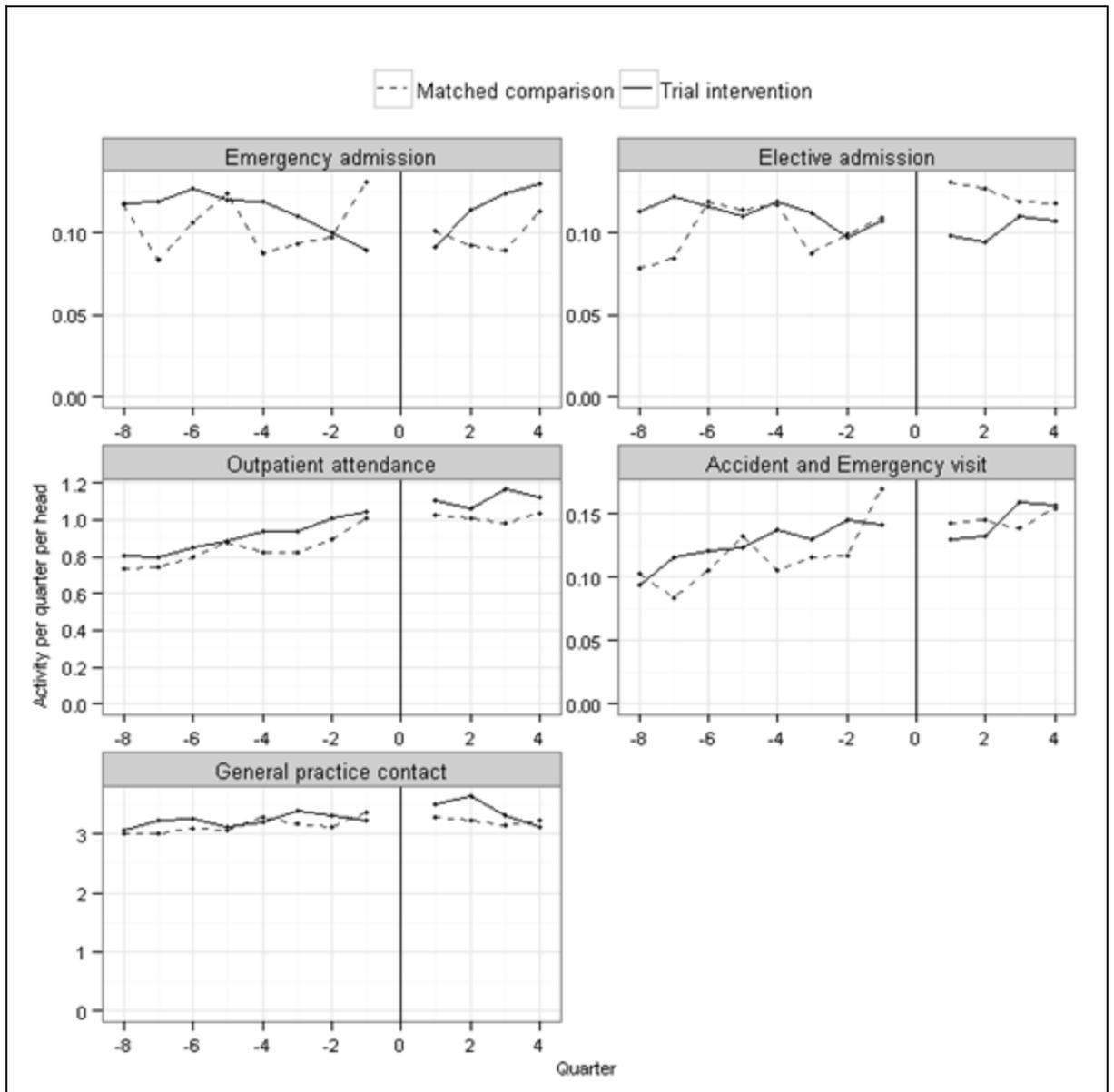
⁶⁹ n=65,552, 841, and 860, for the three groups, respectively.

Table 29: Balance, before and after matching, in the sensitivity analysis as applied to the RCT intervention group (person-level variables, continued)

	Non-participants (n=88,830)	Trial intervention patients (n=1,229)	Matched non- participants (n=1,293)	Standardised difference (variance ratio)	
				Before matching	After matching
Health conditions recorded on hospital data					
COPD	8.0	26.4	23.0	50.1	7.7
Congestive heart failure	5.7	15.1	15.4	31.2	-0.7
Diabetes	24.2	27.2	28.2	6.9	-2.2
Cancer	4.9	6.8	6.5	7.9	1.0
Cancer (benign)	1.4	2.2	1.9	5.9	2.3
Alcohol abuse	0.8	1.4	1.1	5.6	3.0
Hypertension	22.7	37.7	38.5	33.0	-1.7
Injury	8.3	13.8	12.9	17.3	2.6
Iatrogenic	2.9	5.7	5.2	13.9	2.2
Falls	3.4	5.2	5.1	8.9	0.4
Mental health	3.6	4.9	6.0	6.4	-5.0
Angina	5.7	13.1	12.6	25.5	1.5
Ischemic heart disease	10.9	25.1	24.2	37.9	2.1
Asthma	5.8	12.6	12.9	23.8	-0.7
Anaemia	3.5	6.4	6.6	13.7	-0.7
Atrial fibrillation	6.9	16.4	13.9	29.7	6.8
Cerebrovascular disease	3.3	6.0	5.9	12.8	0.7
Peripheral vascular disease	3.1	7.1	6.0	18.4	4.3
Renal failure	2.7	5.6	3.9	14.6	8.0
Respiratory infection	2.8	7.9	6.3	22.6	6.0
Number of chronic conditions per head (mean (SD))	0.90 (1.34)	1.72 (1.78)	1.67 (1.73)	52.1 (1.75)	3.0 (1.06)
Health conditions recorded on primary care data					
Atrial fibrillation	8.4	17.9	13.6	28.3	11.9
Cancer	7.2	9.4	9.2	8.0	0.6
Coronary heart disease	19.7	35.0	34.3	34.9	1.4
Chronic kidney disease	17.6	22.9	22.6	13.2	0.6
COPD	22.0	51.8	47.6	65.0	8.5
Dementia	1.5	0.3	1.8	-12.0	-14.4
Depression	14.8	17.8	19.9	8.2	-5.4
Diabetes	67.8	43.4	44.3	-50.6	-1.6
Heart failure	10.0	29.9	27.8	51.4	4.5
Hypertension	52.7	52.7	54.2	0.0	-2.9
Mental health	1.6	0.2	2.0	-14.1	-16.4
Asthma	13.8	19.8	19.0	16.0	2.1
Stroke / TIA	7.1	9.5	9.4	8.8	0.3

Notes: Data show percentages unless otherwise specified. TIA = Transient Ischaemic Attack.

Figure 16: Crude trends in health service use for the sensitivity analysis



Shows mean numbers of contact per patient per quarter. The observations to the left of the vertical line show rates of primary and secondary care contact for each of the eight calendar quarters preceding trial recruitment (*i.e.*, for a total of two years). The observations to the right of the vertical line show rates for survivors for the four quarters in the trial period. A gap has been imposed at the time of recruitment for clarity. Rates for RCT intervention patients are shown in the solid line. The dashed line shows rates for the subgroup of non-participants who were matched to the RCT intervention patients.

Table 30: Results of the sensitivity analysis (comparison between generalised linear modelling and time series analysis)

	Generalised linear modelling		Time series	
	Comparison with RCT control group	Comparison with non-participants	Comparison with RCT control group	Comparison with non-participants
Emergency admissions per head	0.90 (0.77, 1.05)	1.12 (0.95, 1.31)	0.77 (0.66, 0.91)	1.10 (0.92, 1.31)
Elective admissions per head	0.95 (0.80, 1.14)	0.87 (0.73, 1.05)	0.72 (0.58, 0.90)	0.75 (0.58, 0.98)
Outpatient attendances per head	1.02 (0.93, 1.12)	1.04 (0.95, 1.14)	1.00 (0.92, 1.10)	1.02 (0.93, 1.12)
Emergency room visits per head	0.86 (0.74, 0.99)	0.96 (0.83, 1.11)	0.66 (0.56, 0.78)	0.95 (0.80, 1.13)
Primary care contacts per head	1.06 (1.01, 1.13)	1.04 (0.99, 1.09)	1.09 (1.04, 1.16)	1.06 (1.01, 1.11)
Mortality	0.41 (0.13, 1.23)	1.50 (0.57, 3.94)	n/a	n/a

Notes: Figures are for the incidence rate ratio for the trial intervention group, except for mortality, where figures are for the odds ratio. Time series models were not defined for mortality, as no deaths occurred before the date of enrolment into the trial.

Chapter 7 Future directions for research on telehealth (research paper 5)

7.1 Preamble to research paper 5

The findings of the WSD evaluation presented in Chapter 5 led to a Government initiative to spread telehealth and similar technologies to three million people in England. However, this study did not lend itself to an easy interpretation, especially in light of the assessment of generalisability outlined in the previous chapter. Meanwhile, other research using data collected for the WSD trial has found no impacts on patient-reported outcomes (Cartwright et al. 2013), no overall cost effectiveness (Henderson et al. 2013) and no impacts on general practice contacts (Bardsley, Steventon, and Doll 2013) associated with telehealth. Modest impacts were detected on glycaemic control among a subset of participants with diabetes (Steventon et al. 2014).

The next paper discusses the possible policy responses to WSD and other recently published telehealth studies. It argues that approaches to the evaluation of complex interventions need to be improved, and that there is value in formative studies that monitor estimated treatment effects on a regular basis and refine the interventions over time.

RESEARCH PAPER COVER SHEET

PLEASE NOTE THAT A COVER SHEET MUST BE COMPLETED FOR EACH RESEARCH PAPER INCLUDED IN A THESIS.

SECTION A – Student Details

Student	Adam Steventon
Principal Supervisor	Professor Richard Grieve
Thesis Title	Evaluating complex interventions using routinely collected data: Methods to improve the validity of randomised controlled trials and observational studies

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	n/a		
Have you retained the copyright for the work?*	No	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?	Journal of Telemedicine and Telecare
Please list the paper's authors in the intended authorship order:	Adam Steventon, Martin Bardsley, Brian McKinstry, Richard Grieve, Nicholas Mays, Nicholas Barber, and Harlan M. Krumholz
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)	I conducted most of the background research and prepared the first draft of the manuscript. Co-authors then contributed thoughts and reflections, which I reflected in the manuscript.
--	--

Student Signature: _____

Date: _____

Supervisor Signature: _____

Date: _____

7.2 Introduction

Telehealth has been seen as a promising way to meet the ‘triple aim’ of better care, better health and lower costs for the growing number of people living with long-term health conditions (McLean, Protti, and Sheikh 2011; Berwick, Nolan, and Whittington 2008). The hope is that the technology will prompt more effective oversight and management of these conditions by professionals and patients, thereby improving outcomes while preventing avoidable and costly hospital admissions.

Recent randomised controlled trials have generally not found that telehealth improves outcomes or reduces costs, yet policy initiatives to spread telehealth continue apace. In England, the Department of Health launched a campaign to spread these and similar approaches to 3 million people by 2017 (Department of Health 2012), and incentive payments for providers are now linked to telehealth deployment (NHS Commissioning Board 2013). The apparent mismatch between the evidence and policy direction is explained partly by the complexity of telehealth services. Telehealth has many components, changes over time, and its effects depend on the context into which it is introduced. Therefore, even though some trials have been negative, it can be argued that telehealth might be beneficial in another context or if designed or implemented differently.

Although telehealth is directed at some of the most pressing problems in health services, there is also the potential for wasted resources and adverse events (Takahashi et al. 2010), while some important aspects of care might be going unaddressed. To move forward in a rational manner, it is important to improve the way in which we obtain information about how services such as telehealth work and about their effects.

7.3 Research evidence

Telehealth is part of a broad class of ‘telemedical’ interventions that use technology to deliver health care at a distance. Terminology is not always standardised, but often the distinguishing feature of telehealth services is the use of technology to make regular transfers of medical information (such as blood pressure or symptoms) from patients to health care professionals. In some cases, the medical information is accessed during remote consultations (for example, by videoconference); in others, professionals assess data outside consultations and use rules to determine when to contact patients (for example, if telemonitored blood pressure is over a certain threshold).

There are many ways in which medical information can be incorporated into the management of long-term health conditions, while the settings in which the services operate also vary significantly. Therefore, telehealth is very heterogeneous. Aims also vary, but might include:

- Enabling faster, more appropriate clinician response to deterioration in health (*e.g.*, timely advice about using steroids to manage exacerbations among patients with chronic obstructive pulmonary disease);
- Motivating behavioural change (*e.g.*, by reinforcing the link between health-related behaviour and wellbeing);
- Supporting transitions (*e.g.*, from hospital to the community);
- Supporting diagnosis (*e.g.*, ambulatory blood pressure monitoring for patients with hypertension);
- Providing more accessible support to patients who live in rural or isolated areas; or
- Supporting more coordinated care.

As these interventions are quite different and can be implemented in different ways and contexts, it is reasonable to expect them to have different outcomes.

For a long time, the evidence about telehealth was dominated by small studies of generally poor quality (Bergmo 2009). More recently, some large randomised controlled trials have reported. The largest one so far (the Whole Systems Demonstrator trial) recruited patients with chronic obstructive pulmonary disease, heart failure or diabetes, and found some evidence that patients receiving telehealth had fewer emergency admissions and fewer deaths than patients receiving usual care (Steventon et al. 2012a). However, the interpretation of these results was complicated by the increase in admissions that was observed to occur among usual care participants shortly after their recruitment. This raised questions about whether the reported ‘effect’ on admissions was due to telehealth or to the implementation of the trial protocol among the control group. Telehealth was not found to have an effect on patient-reported outcomes such as quality of life in this trial (Cartwright et al. 2013).

Other large, recent trials have found no change in time to hospitalisation and death among heart failure patients (Chaudhry et al. 2010); and no impacts on clinical outcomes, patient-reported outcomes or costs among people with chronic obstructive pulmonary disease (Pinnock et al. 2013). One trial found that telehealth was associated with lower blood pressure among hypertensive patients, but higher primary care costs (McKinstry et al. 2013). An older trial found that telehealth was associated with tighter glycaemic control in diabetes, but no reductions in costs (Shea et al. 2009).

Previous meta-analyses, which ranged from equivocal to promising, can now be updated (McLean et al. 2011; Baron, McBain, and Newman 2012; Polisena et al. 2009; Chaudhry et al. 2007a; Clark et al. 2007). However, it seems that few, if any, large telehealth trials have conclusively demonstrated simultaneous progress towards better care, better health and lower costs. The most positive news appears to relate to surrogate outcomes such as blood pressure control, but these are often only steps on the way to improved patient outcomes.

7.4 Limitations of the research evidence

One response to the recent trial results would be simply to avoid implementing telehealth, and instead improve health care in other ways, where there is demonstrable evidence of cost effectiveness. When making recommendations about new medicines, technologies and procedures in England and Wales, the National Institute of Health and Care Excellence tends to apply a threshold of £20,000 to £30,000 for each Quality-Adjusted Life Year (QALY) gained. With the Whole Systems Demonstrator reporting a ratio of £92,000 per QALY (Henderson et al. 2013), it is certainly possible to argue that telehealth is not cost effective and should not be introduced.

The alternative view is that telehealth represents a complex change to the way services are delivered with extensive applications and potential, and that many of the trials were undertaken while the approach and technology were relatively new. The impact of telehealth is likely to depend not only on the technology, but also on the whole system, including the interaction between people and the technology. The absence of the desired results in some studies might reflect a lack of understanding about how best to incorporate telehealth into the support of people living with long-term conditions, as well as implementation failure. Consideration of the personal, organisation, cultural, ethical and legal implications of telehealth may have lagged behind technological development (Greenhalgh et al. 2012). Technical improvements might be made in some areas. For example, just as predictive models have been developed to identify patients at high levels of clinical risk (Adams and Leveson 2012), it might be possible to devise smarter algorithms to determine when the telemonitored data necessitate a clinician response. Given the limitations of the available evidence, there is a danger that we might not realise the potential of these approaches because we are being too restrictive too early in what we class as success.

The challenge is how to keep exploring these approaches without wasting resources. Based on our experience, we suggest a focus on three areas.

1) How the technology supports users in their day-to-day lives

Qualitative work has found that patients use telehealth in a variety of, often unexpected, ways (Wherton et al. 2012). Also, patients have given a diverse range of reasons for saying ‘no’ to participation in telehealth trials, including: perceived problems with the technology; preferences for

maintaining existing services; and the perception that telehealth might not support their existing self-care strategies (Sanders et al. 2012). Refusal rates vary significantly in telehealth trials, from 25% to almost 80% (Pinnock et al. 2013; Steventon et al. 2012a). Although these might reflect the perceived burdens of participating in trials, they suggest that telehealth might not always fit in well with the problems that potential users face.

Some work is underway to develop more tailored interventions. For example, ethnographic work is using ‘cultural probes’ to develop better understandings of problems faced by people living with illness, which go beyond the specific medical needs addressed by telehealth devices (Wherton et al. 2012).

2) Developing the underlying logic models

Although the ‘triple aim’ is often cited, the specific aims and objectives of a telehealth service, together with the mechanism by which it is believed to have the intended effects, is rarely stated (Car, Huckvale, and Hermens 2012). For example, although many services have as their ultimate goal improvements in patient outcomes (and, often, reductions in hospital admissions), it is not clear whether these are to be achieved through faster response to deteriorations in health, improved self-care behaviours, or some other means. Ideally, a ‘logic model’ should be devised at the outset that links the specific actions being undertaken to their objectives, with input from the intended users and their carers.

The process of discussing a logic model and writing it down can help reach a better consensus about what should be implemented as part of a telehealth service and how. Logic models have been used to structure evaluation, by informing choices about intermediate outcomes. For example, an intervention aiming to reduce hospital admissions by improving self-care behaviours might examine the impact on those behaviours as an intermediate outcome (Hibbard and Greene 2013), while another intervention predominately focussed on facilitating physician response might track consultation rates or waiting times.

Some studies have found improvements in self-care behaviours (Fairbrother et al. 2013), but not reductions in hospital admissions (Pinnock et al. 2013). Thus in some cases the logic model might need to be updated in response to learning about how telehealth and the surrounding system is functioning, in order to bring about the full range of desired outcomes.

3) A formative approach to evaluation while maintaining rigour

A particular telehealth service might include: rules concerning selection and recruitment of patients; education and self-care support; collection and transfer of medical information; responses to this information; and communication between health care professionals. Success will depend on the

degree of integration with existing services, and on the surrounding context, such as financial incentives, as this shapes how the technology is used (Dixon-Woods et al. 2013). With this in mind, it seems unrealistic to expect that any telehealth service can be fully effective from day one.

As some iteration may be required to develop a satisfactory telehealth service, including improving the technology itself, it is important that evaluation supports efforts to improve services over time. Many traditional evaluation studies are modelled on drug trials, and often take the form of randomising patients into two or more treatments, which are pre-defined. Although these studies are very robust in some important ways, they often assume that the intervention remains broadly the same wherever it is implemented and can be neatly separated from the context in which it is delivered. Unfortunately, randomised controlled trials of telehealth have limited generalisability, as the results are affected by the context and manner of implementation. For example, telehealth may be more effective when integrated into clinical care pathways, but some studies have reported that trial protocols have limited integration (Hendy et al. 2012).

If, within a particular country or locality, policy makers have already decided that telehealth should be introduced, then research efforts might be better targeted at supporting improvement, rather than merely assessing the effects. These more ‘formative’ studies should:

- Be tailored to the local context;
- Provide information on metrics that are tied to local goals and the logic model;
- Provide close to real-time feedback on effectiveness and costs; and
- Be tied to a quality improvement initiative that is able to act on the information about effectiveness. This might mean refining, spreading or stopping telehealth.

Although a series of RCTs could be used for these more formative studies, in many cases the evaluation methods will be nonrandomised. This does not remove the need to attribute any changes in outcomes to specific programmes or activities.

7.5 Developing better evaluations

One potential way forward is to exploit the relatively rich sources of data available to people implementing telehealth, including the large amount of medical information collected by the telehealth systems. In addition, telehealth triage centres hold patient identifiers, and thus data on telehealth users can be linked pseudonymously to routine data sets, such as hospital administrative data or the electronic medical record. The frequency with which routine data sets are updated means that it is possible to track the impact of services in close to real time, the cornerstone of many quality improvement approaches. However, there are problems associated with these data sources that must be addressed (Steventon 2013).

People receiving telehealth in routine clinical practice are likely to be very different from those in a comparison group (*e.g.*, those receiving usual care or another intervention). Thus, unadjusted comparisons between the outcomes of these two groups will give a biased picture of the effect of telehealth. To date, many published nonrandomised studies of telehealth have attempted to sidestep this issue by comparing outcomes before and after the introduction of telehealth for the same group of people (Darkins et al. 2008). Unfortunately, before-and-after studies can overstate treatment effects due to regression to the mean. This arises because telehealth services and their associated evaluations often enrol patients after a recent hospital admission; such patients tend to show reductions in hospital use over time even in the absence of specific new interventions (Roland et al. 2005).

There are several ways to improve on before-and-after designs. One option is to select control patients retrospectively on the basis of them having similar characteristics to those patients receiving telehealth (Chumbler et al. 2008). If these matched control groups are selected properly, then a comparison of their outcomes with those of telehealth patients should reflect the relative effect of the services received, rather than differences in patient characteristics. However, routine data sources such as hospital administrative data might not contain sufficiently detailed information to adjust for important differences in the characteristics of telehealth and non-telehealth patients (Rubin 2010).

Matched controls are being used in telehealth research, with promising findings (Barnett et al. 2006; Baker et al. 2011). However, since it is not clear to what extent routinely collected data contain sufficiently detailed information on patient characteristics, more methodological work is needed in parallel with greater uptake of these methods. This could encompass sensitivity analyses that illustrate the degree of robustness to unobserved confounding (Keele 2010). Newer methods such as ‘synthetic controls’ are being developed to compare a region’s outcomes with a weighted average of the outcomes of several control regions (Abadie, Diamond, and Hainmueller 2010).

If done appropriately, extending data linkage to more data sets would help deal with confounding, as well as extend the range of outcomes that can be assessed. Feedback to local health practitioners also needs to be made faster in order to create a learning system.

7.6 More systematic data collection

In practical terms, there are a number of ways in which information systems could be configured to support more timely evaluation in this area. At a local level, the funders of care might use local data sets to conduct analysis. More progressive telehealth providers might also start to integrate controlled evaluation methods into their normal practice, in order to improve their services and demonstrate value to their customers. At a regional or national level, it is possible to envisage initiatives that would invest in infrastructure to systematically make routinely collected data from telehealth devices

available for research across whole health systems. Rather like the way disease registries have been used, there are significant advantages in pooling intelligence across suppliers and settings. A similar registry for telehealth would enable research to be conducted at greater scale in a way that can explore the impact of context on outcomes.

There is another reason to improve data collection. If those implementing telehealth services learn from continuing evaluation and respond with more targeted, useable and integrated solutions, then these services might be more effective than those of the past. However, to go too far down the wrong track risks embedding ineffective interventions into health care. Therefore, we need to review the spread of these technologies. This would seem particularly valuable in England where, despite the stated ambition to reach '3 million lives', it is not clear how progress towards this goal is being assessed. For the time being, there appears to be some distance to go before telehealth produces the sought-after improvements in patient outcomes and efficiency, and those implementing telehealth need to be prepared to develop and refine their approaches.

Box 8: Key messages

- Recent RCTs have not conclusively demonstrated the effectiveness of telehealth in improving care, improving health and lowering costs. Yet, there might be opportunity to increase its effectiveness by a clearer focus on how the technology is actually used and the mechanism by which it is intended to have an effect.
- As a telehealth service consists of several different components, it seems unrealistic to expect that it can ever be fully effective from day one or work in the same way across all settings. A formative approach would involve monitoring outcomes on a regular basis and refining the service over time.
- Studies that track the outcomes of patients receiving telehealth in routine clinical practice in close to real-time appear to be crucial to support a learning system. These studies could exploit routinely collected data and use matched controls, though more methodological development is needed in parallel, in particular to address inevitable concerns about confounding.
- More systematic data collection concerning patients receiving telehealth in routine clinical practice would help measure the spread, cost and effectiveness of telehealth.

Chapter 8 Discussion

8.1 Introduction

This thesis addresses gaps in the methodological and applied literatures pertaining to an area of high-policy relevance, namely the evaluation of complex interventions for people living with long-term health conditions. My focus was on the estimation of treatment effects in these studies, particularly the internal validity of observational studies and the generalisability of RCTs. Using routinely collected data throughout, I had the following specific objectives:

1. To critically review methods appropriate to addressing the internal validity of observational studies of complex interventions and the external validity of RCTs of these interventions (Chapter 2).
2. To estimate the effect of Birmingham OwnHealth on the utilisation of hospitals, by employing an observational study design (Chapter 3).
3. To assess the implications of the choice of control population in observational studies for the bias and efficiency of the resulting estimates of relative treatment effect (Chapter 4).
4. To estimate the effect of telehealth on hospital utilisation and mortality, within a large RCT (Chapter 5).
5. To develop an analytical method to assess the generalisability of RCTs, together with a sensitivity analysis to explore the consequences of potential causes of non-generalisability (Chapter 6).
6. To make recommendations for policy development and evaluation methods (Chapter 7).

The next section briefly reviews the methodological gaps identified for both observational studies and RCTs. The findings, contributions and limitations of my work are then presented separately for observational studies and RCTs in Sections 8.3 and 8.4. Section 8.5 summarises the recommendations for policymaking and evaluation methods, while Section 8.6 outlines some areas for future research. The final section concludes.

8.2 Findings from the review of the methodological literature

8.2.1 *Improving the internal validity of observational studies*

While there are many methods for dealing with confounding in observational studies, the critical literature review in Chapter 2 identified matched control methods as being particularly suitable for the current context. These benefit from the availability of a clear diagnostic (*i.e.*, balance) (Hill 2008). Moreover, if the data are first pre-processed using a matching algorithm, this can reduce the

dependence of the estimated treatment effects on model specification versus applying regression methods to unmatched data (Ho et al. 2007). Unobserved confounding will be a concern in matched control studies, and strategies to deal with it require good evaluation design (Rubin 2010). Studies have compared strategies for selecting the variables included in the matching algorithm (Austin, Grootendorst, and Anderson 2007), but one aspect of study design that has received scant attention in the methodological literature concerns the choice of higher-level unit(s) from which to select the matched control patients. This thesis therefore examined the impact of the choice of control population on balance within the context of a particular case study, as well as the implications for bias and mean-squared error through simulations.

8.2.2 Assessing and extending the generalisability of RCTs

Generalisability is widely recognised as a concern in RCTs, and it may explain why technologies that have been shown to be beneficial in RCTs have often not been diffused into routine clinical practice (Rothwell 2005). Much of the effort that has gone into improving the generalisability of RCTs has focussed on design-stage strategies, such as the use of pragmatic trial designs (Roland and Torgerson 1998). Analytical approaches to assessing generalisability have until recently been limited, for example relying on interpreting the 'representativeness' of patients from tables of baseline characteristics (Schulz, Altman, and Moher 2010).

Building on earlier work (Stuart et al. 2011), recent advances in the analytical literature include the use of placebo tests to establish the feasibility of adjusting RCT-based estimates of the sample treatment effect to produce unbiased estimates of population treatment effects (Hartman et al. 2015). However, these tests have only been applied to a relatively narrowly defined surgical intervention, namely pulmonary artery catheterisation, and not to more complex interventions. Also, Hartman and colleagues focussed on the adequacy of the methods used to adjust for differences in patient characteristics between settings, and they did not consider whether the intervention and control regimens tested within the RCT setting were the same as those delivered in routine practice. This is an important issue for RCTs of complex interventions, as the implementation of the trial protocol might affect the nature of the interventions being delivered in unintended ways (Chapter 1). Furthermore, previous work has not addressed what can be done when the placebo tests fail. Therefore, the thesis developed placebo tests for RCTs of complex interventions, and proposed a new form of sensitivity analysis to assess the consequences of potential causes of non-generalisability.

8.3 Discussion of objectives relating to observational studies

8.3.1 Applied case study - evaluation of telephone health coaching service

As well as the methodological contributions outlined above, the thesis aimed to add to the applied literatures by evaluating specific complex interventions. Observational, matched control methods

were used to evaluate a large telephone-health coaching service for people with long-term health conditions in a large English city (namely, Birmingham OwnHealth, see Chapter 3). Patients enrolled into this service received a series of telephone calls from nurses and a personalised care plan. In conducting this study, I retrospectively linked health-coached patients to national hospital administrative data and selected matched control patients using prognostic score matching ($n=2,698$ patients in each group). Matched controls were sourced from areas of England that were similar to the area that offered the intervention, rather than from within the intervention area. This was partly because, at the time of the study, it was unclear which strategy to control area selection was optimal. Moreover, there were concerns about spillover effects and non-identification of intervention patients from within the administrative data.

Mortality rates were similar between the health-coached patients and matched controls, while the health-coached patients experienced a faster increase in emergency hospital admissions than matched controls (difference 0.05 admissions per head over 12 months, 95% confidence interval 0.00 to 0.09, $p=0.046$). Because confounding remained a concern, I conducted sensitivity analysis to assess the potential impact of hypothetical unobserved confounders on the findings, using a simulation method (Higashi et al. 2005). This revealed that, in order for the finding of a reduction in emergency hospital admissions to have been reversed, the person-level unobserved confounder would need to be more strongly correlated with treatment assignment and outcome than was realistic based on the previous literature (Jordan, Lancashire, and Adab 2011).

The study was believed to be the largest observational study of telephone health coaching at the time of publication. The findings were consistent with a previous observational study and a previous RCT, both of which concerned low-income or disabled adults in the U.S. Medicaid programme (Lin et al. 2012; Kim et al. 2013).⁷⁰ However, a separate RCT of telephone health coaching for privately-insured individuals, also in the United States, had found reductions in hospital utilisation (Wennberg et al. 2010). Discrepancies with my study might have been due to the use of predictive risk models to identify cohorts of patients for health coaching in the RCT, or to their inclusion of an element of shared decision making for preference-sensitive conditions, as discussed in Chapter 3. Since my study was published, another study has reported findings for the impact of Birmingham OwnHealth on hospital admissions (Nymark et al. 2013). This study found fewer hospital admissions among health-coached patients than matched controls over the year following enrolment (0.61 *vs.* 0.84 per head, a relative reduction of 27%). These authors matched on some clinical metrics that were not available in my data, including HbA1c, but they did not report balance on the prior rate of hospital admissions, which is arguably the most important prognostic variable (Billings et al. 2013). Another matched control study found that telephone health coaching was associated with lower health service

⁷⁰ Medicaid is a publicly funded health insurance programme for families and individuals with low income and limited resources in the United States.

expenditures than usual care, but not reductions in emergency department visits (Jonk et al. 2015). A further RCT is underway, this time in Germany, but has not yet reported (Dwinger et al. 2013).

Overall, the study in Chapter 3 demonstrated an approach to evaluation based on linking patients to national hospital administrative data, forming matched control groups, and conducting sensitivity analysis for unobserved confounding. Although the sensitivity analysis suggested it was implausible that unobserved confounding at the individual level obscured reductions in hospital admissions, a residual concern related to control area selection. Confounding could have been introduced at the area level as a result of selecting matched control patients from areas similar to Birmingham, rather than from within Birmingham itself.

8.3.2 Implications of the choice of control population

To explore issues around the choice of control population, Chapter 4 began by reanalysing another matched control study, this time a rapid-response service that was delivered as part of the Partnership for Older People Projects (POPPs) (Steventon et al. 2012b). In the original evaluation, matched controls had been selected from areas of England that were matched to the characteristics of the intervention area, in a similar manner to the Birmingham OwnHealth study. My reanalysis investigated three further approaches to control area selection, namely local controls, controls from arbitrarily chosen areas of England, and national controls.

In the reanalysis, lower standardised differences were achieved on observed, person-level variables when using external rather than local controls. To assess the potential effect of the choice of control population on the balance of unobserved variables, two important prognostic variables were omitted from the matching algorithm, namely age and the predictive risk score. Imbalances on these variables were also minimised when using external rather than local controls. Treatment effects varied markedly depending on the choice of control area, but in the case study the variation was minimal after adjusting for the characteristics of areas.

Simulations were used to assess the implications of the choice of control population on the bias and mean-squared error of the resulting treatment effects, as this was not possible within a case study setting. The simulations found that selecting controls from a matched area produced the least bias of all the strategies, provided that two restrictive assumptions were met. First, area-level variation in outcomes must either not exist or be largely explained by area-level variables accounted for in the matching. Second, unobserved confounding must be a threat within the local area, for example because there is a strong confounder and high intervention saturation. The study concluded that, when the endpoint relates to hospital admissions in England, there is likely to be a large degree of unexplained variation in the outcome between geographic areas. The impact of this variation dwarfed

the benefits of the better balance achieved on individual-level variables when using external controls, leading us to prefer local controls for analyses of this endpoint.

8.3.3 Contributions of this thesis for the design of observational studies

Careful design is of paramount importance to observational studies since, however advanced the analytical method, the study is likely to be biased if the underlying assumptions are not met (Rubin 2008). Investigators have used a range of approaches to define the control population when evaluating health care interventions, but the relative benefits of some popular design choices (in particular, local or external control groups) have rarely been directly assessed. Although within-study comparisons and meta-epidemiological analyses have been done, these can rarely account for all confounders of study design (Deeks et al. 2003). The few studies that have assessed control area selection directly have either focused on methods that use several control groups (Rosenbaum 1987; Lu and Rosenbaum 2004) or have assumed no unobserved confounding at the individual level (Stuart and Rubin 2008). The former approach has been criticised because it does not provide information about which control group produces least bias (Rothman, Greenland, and Lash 2008), while the latter approach does not address a critical threat to the validity of observational studies (Rubin 2010).

The papers contained in this thesis will help inform the design of future observational studies. For example, the interim evaluation of the Pioneer Accountable Care Organisations in the United States reported treatment effects estimated using two sets of controls, one selected from within the local market for health care and another from similar markets elsewhere (L&M Policy Research LLC 2013). In this study, there was little indication about which control group should be preferred, but future studies could be informed by the conclusions in Chapter 4. One practical message is that, if there is likely to be substantial unexplained variation in outcomes between areas, a local approach to control groups should be preferred, even if the intervention saturation is high and the availability of local controls is limited. If these research findings had been available when designing the Birmingham OwnHealth study, they would have provided greater motivation to use local controls, notwithstanding concerns about spillover effects and non-identification of intervention patients. This could have altered the estimated treatment effect in either direction.

8.3.4 Limitations in relation to observational studies

While the simulations were calibrated to case study data, no simulation study can consider every setting that might be relevant in applied research. Observed variables were relatively easy to balance in all of the scenarios considered, but in the more general situation other matching methods, such as genetic matching, might be required to maximise covariate balance (Diamond and Sekhon 2013). The simulations did not address situations in which adequate balance on observed variables was simply not possible within the local area, even when more sophisticated matching algorithms are used. In these instances, it is likely that the data are not adequate to answer the causal questions being asked of

them (Rubin 2010), as was the conclusion in a recent study examining Virtual Wards (Lewis et al. 2013).

For simplicity, the simulations did not address the possibility of selecting multiple matched control areas for a single intervention area, as was the approach taken in the Birmingham OwnHealth study (Chapter 3). However, the existing literature about the relative performance of 1-1 versus 1-many matching in a single-level setting is informative (Austin 2010). It is reasonable to suppose that using multiple matched areas would reduce the variability of treatment effect estimates compared with strategies that use a single matched area, but that it would lead to higher levels of covariate imbalance, and so increase bias.

The simulations assumed a single intervention area, but other strategies become available in more general settings. For example, Griswold and Localio (2010) fitted a single, multilevel propensity score model using individual-level data from several hospitals. This model aimed to reflect both area-level and individual-level decision-making and it was used within a one-step matching process. In these more general settings, it is possible to select a varying number of matched control areas for each intervention area, a strategy that has been shown to be effective at reducing bias in the single-level setting compared with using a fixed number of controls per intervention patient (Ming and Rosenbaum 2000). The implications of these strategies were not assessed in Chapter 4.

In determining the approach to control area selection, investigators will need to consider factors such as spillover effects, differences in measurement between areas, and data availability. These were not the focus of this thesis. Although the simulation study found substantial variation in hospital admissions between areas of England that could not be accounted for by observed, individual-level variables, it did not set out to identify variables that could explain this phenomenon.

The limitations of the case study are described in Chapter 3 and include the possibility of unobserved confounding and a single-variable approach to sensitivity analysis.

8.4 Discussion of objectives relating to RCTs

8.4.1 Applied case study - evaluation of telehealth

The second case study concerned the Whole Systems Demonstrator (WSD) cluster randomised trial of telehealth compared with usual care (Chapter 5). As selection bias is a potential concern in non-blinded cluster randomised trials, treatment effects were estimated after adjusting for differences in important prognostic variables between intervention and control groups, including differences in a predictive risk score.

The study reported that telehealth was associated with lower emergency hospital admission rates than usual care over 12 months (incidence rate ratio 0.81, 0.65 to 1.00, $p=0.046$), as well as lower mortality

at 12 months (odds ratio 0.54, 0.39 to 0.75, $p < 0.001$). Subsequent studies of data collected for this trial have found no impacts on rates of general practice contact (Bardsley, Steventon, and Doll 2013) or quality of life, anxiety or depressive symptoms (Cartwright et al. 2013), and telehealth was not cost-effective (Henderson et al. 2013). No clear subgroup effects were reported for the primary outcomes (Newman et al. 2014), though modest improvements in glycaemic control were detected among a subset of patients with diabetes (Steventon et al. 2014). Further studies are ongoing (Newman et al. 2014), for example regarding impacts on self-efficacy.

More recent RCTs have reported that telehealth was associated with modest improvements in clinical metrics such as blood pressure (Stoddart et al. 2013). However, they have not replicated the lower rates of hospital use reported by WSD (Vuorinen et al. 2014; Pinnock et al. 2013). The U.S. Veterans Health Administration has reported an updated observational study of the effectiveness of its large telehealth deployment (Darkins et al. 2015, 2008), this time assessed against a matched control group. Like the previous study, the updated study reported that telehealth was associated with fewer hospital admissions than usual care. Compared to other studies including WSD, the telehealth service in the Veterans Health Administration was more tightly specified and was also targeted at different groups of patients, with a particular focus on people with mental health conditions. Updated systematic reviews are needed as, although several reviews have concluded that telehealth may reduce admissions (McLean et al. 2013), these have not reflected the most recent trials.

Chapter 5 examined the potential limitations of the WSD study with respect to both internal and external validity. Tables of baseline differences appeared to confirm that randomisation produced similar treatment groups (*i.e.*, estimates were internally valid under the intention-to-treat method applied), despite some differences in the median number of patients per cluster. Of greater concern were the patterns of emergency admissions for the control group, which were observed to increase shortly after their recruitment. The paper discussed three possible explanations for this increase. First, it could reflect the natural pattern of admissions for these patients. If this were the case, then generalisability would not have been compromised. Alternatively, the health care professionals who recruited patients into the trial might have identified unmet need during the recruitment process, resulting in additional health care being provided to control patients. This explanation assumes that health care professionals were content to use telehealth to manage any additional need identified within the intervention group. Finally, it is possible that some patients in the control group reacted to their treatment allocation in ways that meant they sought more care from emergency departments. This theory, if true, would imply that the trial had poor generalisability with respect to emergency department utilisation, as the comparison was not with 'usual care' as defined in the target population. Thus, the case study illustrated the need for an empirical approach to assess the

generalisability of RCTs. This is particularly important as the paper has been widely cited,⁷¹ and led to an initiative to increase the numbers of people receiving telehealth and similar assistive technologies in England to three million (Department of Health 2012), though this policy has since been revised.

8.4.2 Placebo tests and associated sensitivity analyses

This thesis proposes an approach to assessing aspects of generalisability, based on using placebo tests to compare the outcomes of trial participants and non-participants receiving similar treatments, after adjusting for differences in their baseline characteristics. Ideally, investigators would be able to confirm that these adjusted outcomes are similar. In these instances, policymakers could have confidence that, first, the treatments were similar across settings and, second, important differences in the characteristics of patients had been identified (assuming multiple effects did not cancel out).

The proposed approach to placebo tests has several important features. First, it recognises that generalisability relates not only to the characteristics of the research participants but also to the treatments themselves. This is particularly important when evaluating complex interventions because these interventions can interact with their context, as described in Chapter 1. Second, the placebo tests harness routinely collected data, meaning that they can be readily applied to RCTs that use these data to define their endpoints and eligibility criteria. Third, in the event that the placebo tests fail, sensitivity analyses are proposed to examine the consequences of potential causes of non-generalisability.

When placebo tests were applied to the WSD data (Chapter 6), the placebo tests passed or were close to passing for rates of general practice contacts, outpatient attendances and elective admissions. Thus, we can be confident that the RCT control group used these types of health care at a similar rate to matched non-participants. However, the RCT control group experienced more emergency hospital admissions and deaths than the matched non-participants, and the placebo test failed for these two endpoints. Although the estimate for either endpoint might have been biased due to unobserved confounding, these findings supported the theory that some patients in the control group reacted to their treatment allocation in ways that meant they sought more care from emergency departments. Regardless of the reason for the failure of the placebo tests for emergency admissions and mortality, the conclusion is the same: there was no evidence that the trial was generalisable with respect to these two endpoints.

The sensitivity analysis for emergency admissions showed that, if one assumes that the placebo test failed due to differences in treatments between settings rather than due to differences in patient characteristics, then telehealth was associated with a trend towards an increase rather than a reduction in emergency hospital admissions.

⁷¹ 194 citations according to Google Scholar, as at 3 March 2015.

8.4.3 Contributions of this thesis for the analysis of RCTs

The proposed placebo tests and sensitivity analyses build on previous work in this area. Many previous comparisons of estimated treatment effects between randomised and nonrandomised studies have aimed to establish whether the method used to address patient-level confounding in the nonrandomised study was reliable (LaLonde 1986; Dehejia and Wahba 2002). I have argued that this approach fails to recognise a crucial aspect of the treatment assignment mechanism, which concerns recruitment into the trial, rather than treatment allocation within routine practice. Also, existing comparisons have not often met the quality criteria that have been developed in this area (Cook, Shadish, and Wong 2008). For example, they have not standardised measurement across randomised and nonrandomised settings (LaLonde 1986). Indeed, a systematic review of comparisons between RCT and non-RCT patients (Gross et al. 2006) found that very few of these made any adjustment for differences in patient characteristics between the two settings. The use of placebo tests is therefore a particularly important innovation (Jones 2007). While another recent study proposed using placebo tests to assess generalisability (Hartman et al. 2015), these authors did not propose a strategy for estimating population treatment effects in settings where the placebo tests failed, whereas in this thesis I proposed combining placebo tests with sensitivity analysis.

One of the contributions of this thesis is that it applies placebo tests to an RCT of a complex intervention that was delivered outside of hospital settings. The outcomes of participants in RCTs of these interventions have not often been compared to those of non-participants (Gross et al. 2006). A recent study in Austria compared outcomes for patients receiving disease management in a cluster randomised trial (n=649) with those of patients receiving similar programmes routinely (Nolte and Hinrichs 2012). Metabolic control was found to improve more quickly amongst trial participants than non-participants after adjusting for differences in patient-level characteristics, but the non-participants were selected from other regions of Austria, and thus regional or provider characteristics might have confounded the results. Moreover, the comparisons were not formulated as placebo tests. In another study, parallel randomised and observational studies were undertaken of an intervention involving remote patient monitoring during nurse teleconferences (Pekmezaris et al. 2012). Both of the constituent studies reported no effect on hospital admissions, but the inclusion criteria were applied more strictly in the randomised than the observational study.

The importance of placebo tests is underscored by a recent RCT of telehealth for patients with multiple morbidities (n=205) (Takahashi et al. 2012). This reported lower mortality rates among RCT control patients than non-participants, after standardising for a risk score based on administrative data (3.9% compared with 13% of patients over 12 months). The RCT controls also experienced lower mortality rates than telehealth patients (3.9% compared with 14.7%). Thus, although the Takahashi study reported a treatment effect for mortality that had the opposite sign to that from

WSD, similar concerns arose from both trials about the possibility that the reported treatment effects were due to unexpected patterns among trial control patients.

8.4.4 Limitations

The study in Chapter 6 only addressed only some aspects of generalisability, namely those related to the characteristics and treatments of the control group. The generalisability of the intervention regimen could have been addressed with very similar methods, but comparable data were not available on patients receiving telehealth in routine practice settings. Moreover, the method cannot determine whether the placebo tests failed due to differences in the treatments or differences in the baseline characteristics of participant and non-participant groups. However, both possibilities imply problems with the generalisability of the RCT.

In the WSD trial, routinely collected data were used as the source of information on outcomes, and moreover the trial inclusion criteria were such that it was possible to identify eligible patients from within these data. However, this will not be the case for all trials, and thus placebo tests may require additional data collection in some instances or might not be feasible. For example, it would be challenging to apply placebo tests to the other RCT that was conducted as part of the WSD programme, which concerned the effect of remote, automatic and passive monitoring for people with social care needs ('telecare'). Although the primary endpoint of this trial also related to hospital admissions (Stevenson et al. 2013a), it was not possible to establish the eligibility of potential comparison patients from within routinely collected data, as these data did not contain sufficient information on social care needs. Where resources are constrained, a question remains about the relative importance of collecting data to assess generalisability versus, for example, extending the trial follow-up period.

In developing the placebo tests, we did not address difficulties associated with multiple endpoints. These difficulties are well illustrated by the WSD trial, for which the placebo tests passed for some endpoints but not others. We suggested that in future studies one could weight the alternative endpoints and apply the placebo tests to overall metrics such as those used in cost-effectiveness analyses (for example, to contrast net monetary benefits between the RCT and routine practice settings). Finally, we did not make extensive comparisons between alternative estimators for the placebo tests, or consider situations with missing endpoint data.

8.5 Recommendations for policymaking and evaluation methods

Health services around the world face a pressing need to improve the quality of care being delivered to people living with long-term health conditions, to improve outcomes for these people, and to reduce costs. One recent initiative in England is the Five Year Forward View, which emphasises the need for new models of care that "break down the barriers in how care is provided between family

doctors and hospitals, between physical and mental health, [and] between health and social care" (NHS England 2014).

The methods developed in this thesis can improve the design and analysis of evaluations of complex interventions such as those proposed by the Five Year Forward View. However, it must be acknowledged that, in keeping with the literature, this thesis did not provide evidence to suggest that any of the interventions considered reduced hospital admissions (Purdy 2010; Dusheiko et al. 2011). Unfortunately, little information is available to understand the reasons for the failure of many interventions to reduce admissions. Possible explanations include: the interventions were targeted at patients unlikely to benefit from them; there were problems with the implementation such that fidelity to the original intervention design was lost; the hypothesized mechanisms for the effect were incorrect; or the contexts in which the interventions operated were not conducive to the changes being proposed.

Chapter 7 argued that, for progress to be made in improving out-of-hospital health care for people with long-term conditions, the evaluation methods themselves need to change to provide information that is more closely tailored to the decisions that must be made by policymakers, practitioners and patients. Evaluations must also be more focussed on how patients interact with the services and acknowledge that services are not fully effective from day one and work in a different way depending on the local context. Moreover, in rolling out the intervention, it is useful to support learning and improvement and continue to evaluate the changing nature of the intervention. These observations motivate some avenues for future research that build on the contributions made by this thesis.

8.6 Future research

This thesis did not seek to determine whether randomised or observational methods are more appropriate for evaluating complex interventions. Although the debate about the relative advantages of these methods will continue, in some ways it is a false dichotomy. Indeed, this thesis has demonstrated a way to combine observational and randomised methods with placebo tests, which preserve the benefits of randomisation while also testing some of the assumptions required for generalisability. Because both observational and randomised studies are needed, efforts should be placed on maximising the validity of both. The following sections briefly outline some research agendas that could be pursued.

8.6.1 The sufficiency condition for the internal validity of observational studies

Both theoretical (Rosenbaum and Rubin 1983; Drake 1993) and empirical work (Shadish, Clark, and Steiner 2008) has showed that observational studies can yield unbiased estimates of treatment effects if the assumptions underlying their analytical methods are met. This thesis focussed on methods to reduce the likelihood of bias from unobserved confounding by means of careful selection of the

control population. However, the question remains about whether routinely collected data are adequate to answer the questions being asked of them (Rubin 2010). Several strands of research could be explored, relating to the treatment assignment mechanism, the data sets, control by systematic variation, and sensitivity analyses.

8.6.1.1 Understanding the treatment assignment mechanism

Rubin's conceptual framework for dealing with confounding in observational studies rests on understanding the treatment assignment mechanism (Chapter 2). The theory underlying many methods from causal inference treats this mechanism as being knowable (Rosenbaum and Rubin 1983), but for complex interventions it is a product of a dynamic and unpredictable system, and so may be unknowable. The approach to placebo tests in Chapter 6 followed best practice for observational comparisons, and variable selection was informed by a qualitative study about the reasons given by patients for declining to participate in the RCT (Sanders et al. 2012). However, there was still the problem about how well the themes emerging from the qualitative study could be mapped onto the routine data sets, and moreover the qualitative study only reflected part of the treatment assignment mechanism, and not the full complexity of how patients, clinicians and others interact with each other over time. An alternative approach that may be more successful at reducing bias within observational studies of complex interventions focuses on modelling the relationships between patient characteristics and outcome, for example using the prognostic score (Hansen 2008a). However, experience suggests that some outcomes can be predicted with greater accuracy using routinely collected data than others.

When placebo tests were applied to the WSD trial in Chapter 6, they failed for mortality in ways that could not easily be explained by differences in treatments, suggesting that unobserved confounding was responsible. Furthermore, some observational studies have reported unexpected differences between the mortality rates of patients receiving complex interventions and those of matched controls (Steventon et al. 2012b; Steventon, Bardsley, and Mays 2014; Roland et al. 2012). It is true that unexpected mortality differences have also been detected in RCTs of complex interventions (Alkema et al. 2007; Takahashi et al. 2012), and they were not present in the observational study of Birmingham OwnHealth presented in Chapter 3. Nevertheless, it is a plausible hypothesis that unobserved confounding is more likely to occur when routine data sets are used to study health outcomes such as mortality than when they are used to study the utilisation of hospital care. Indeed, health is influenced by many more factors than clinicians can realistically take into account when deciding whether to admit patients (Marmot and Wilkinson 2009), suggesting that there are more potential confounders of mortality than admissions. Although reasonably accurate prognostic models have been built for hospital admissions and readmissions (Billings et al. 2006; Tan et al. 2013), models for mortality have tended to focus on immediate post-operative mortality, rather than mortality over,

say, a 12 month period for people with long-term conditions (Aylin, Bottle, and Majeed 2007; Burroughs et al. 2006). Further work is warranted to investigate which outcomes can be predicted most accurately from routinely collected data. Moreover, the performance of prognostic score matching for these outcomes could be compared with that of propensity score matching in terms of the bias and statistical precision of the resulting treatment effects.

8.6.1.2 The value and accuracy of different types of routine data

A distinction can be made between routine data that are generated during the delivery of care (such as the electronic medical record) and those produced by the administrative systems of health care organisations (Aylin, Bottle, and Majeed 2007). One study found that propensity score matching using administrative variables was not sufficient to balance additional variables available from clinical data (Austin et al. 2005). Further work is needed to assess whether this finding holds true in the more general setting, since it has implications for the value of these alternative data sets in observational studies. Moreover, studies could attempt data linkage across sectors or could harness the growing amount of routine data that is patient, rather than clinically, generated (Nelson et al. 2015).

There is also the long-standing question about the accuracy and relevance of routine data (Roos et al. 1993; Spencer and Davies 2012). One area where progress could be easily made is in comparing the accuracy of self-reported and administrative data on the use of services. Several existing comparisons are now quite out-of-date (Wallihan, Stump, and Callahan 1999; Bellón et al. 2000) and did not address social care. I am currently working on a comparison between the routine and survey data collected for the WSD trials. If data from a third source were available, then latent class modelling could be used to estimate the accuracy of data from each source (Dendukuri et al. 2005).

8.6.1.3 Extending control by systematic variation

Another strategy that could be explored exploits the large size of routine data sets to generalise Campbell's 'control by systematic variation' (Campbell 1969). Indeed, Chapter 4 suggests a potential diagnostic criterion for unobserved confounding, namely the degree to which treatment effects vary over many alternative control populations after adjustment for both individual-level and area-level variables. The application of this diagnostic in the current setting may, however, require better measurement of the reasons for area-level variation in outcomes than currently exists (Fisher, Goodman, and Chandra 2008).

8.6.1.4 Extending simulation methods for assessing the potential impact of unobserved confounding

The sensitivity analysis developed for the Birmingham OwnHealth study (Chapter 3) assessed the robustness of the evaluation findings to unobserved confounding. Important aspects of the approach include: simulating hypothetical unobserved confounders with varying levels of association with the treatment assignment and outcome; assessing the strength of association required to overturn the

headline evaluation findings; and assessing whether these associations are likely to have occurred, based on the existing literature. My work developed a method from previous applied work (Higashi et al. 2005) but Jennifer Hill and colleagues have since proposed a more general approach to these sensitivity analyses, which I have applied in a subsequent telehealth study using data from North Yorkshire (both articles are currently under peer review). Potential extensions of Hill and colleagues' method would address binary endpoints or multiple confounders.

8.6.2 Use and development of placebo tests

Following recent developments in the literature about assessing the generalisability of RCTs (Cole and Stuart 2010; Hartman et al. 2015), there is now the opportunity to apply placebo tests to more examples, which may raise further challenges for the methods. Placebo tests may be particularly applicable to pragmatic RCTs that are embedded within routine data sets. These include randomised-registry trials (Lauer and D'Agostino 2013), such as the recent Thrombus Aspiration in ST-Elevation Myocardial Infarction (TASTE) trial in Scandinavia (Fröbert et al. 2013; Lagerqvist et al. 2014). This used national health registry data to assess the long-term mortality outcomes of both RCT participants and non-participants, but did not assess the implications of differences in the baseline characteristics of these groups within placebo tests. Another category of trials that might be well suited to placebo tests is the point-of-care RCT, as described by Van Staa and colleagues. Here, the electronic medical record in primary care is used to alert doctors of a patient's eligibility for the trial during the course of a regular consultation, to record treatment decisions, and to obtain information about outcomes (van Staa et al. 2012). The approach was piloted in the Randomised Evaluations of Accepted Choices in Treatment (REACT) trials for statins and antibiotic treatments (van Staa et al. 2014), but the authors reported problems obtaining research governance approvals and that clinicians chose to apply the method to those patients who did not have an acute illness. Nevertheless, investigators from this and similar trials have been optimistic about the long-term applicability of the method (Mosis et al. 2006; D'Avolio et al. 2012). In general, placebo tests may be appropriate in comprehensive cohort studies, in which patients who do not consent to randomisation are nevertheless followed up, either using routinely collected data or primary data collection (Olschewski, Schumacher, and Davis 1992; Brewin and Bradley 1989).

More routine use of placebo tests would build up valuable evidence about the empirical generalisability of RCTs of complex interventions, and enable more precise meta-epidemiological assessment of the effect of alternative aspects of trial design on the level of generalisability achieved. The WSD studies presented in Chapters 5 and 6 suggest that it is important to understand the effects on outcomes of the information provided to patients during the recruitment stage of an RCT, and likewise the effects of patient preferences for the intervention versus control (Zelen 1990).

Many of the outstanding questions about placebo tests (see Section 8.4.4) concern the relative merits of alternative estimation approaches (*e.g.*, time series versus generalised linear modelling). The relative merits will differ by context and it would be useful to consider them. In situations in which the new treatment of interest is offered routinely as well as within an RCT, then it is possible to compare the outcomes of treated patients between settings. This is the approach that Hartman and colleagues took for pulmonary artery catheterisation (Hartman et al. 2015), and could be taken for telehealth now that the technology is being introduced more widely in England.

8.6.3 Specifying and testing the logic model

Frangakis pointed out the importance of understanding the mechanism by which interventions are believed to have an effect when assessing the generalisability of evidence from RCTs (Frangakis 2009). Although his argument applies to any mediator of the treatment effect, he illustrated the issues using the example of non-compliance, and divided both RCT and routine practice populations into principal strata (*i.e.*, into 'never-takers', 'always-takers', and 'compliers') (Frangakis and Rubin 2002). He showed that the predicted outcomes for people allocated to the treatment in routine practice will in general be biased if the analysis does not reflect differences in the distribution of principal strata between settings (Frangakis 2009).

Chapter 1 highlighted more general reasons why attention should be given to the mechanism by which interventions are believed to have their effect. That discussion focussed on telehealth, and identified several theories from the peer-reviewed literature about why this kind of intervention might reduce hospital admissions, for example by improving self-management skills, improving drug titration, or facilitating coordination between care teams. In other words, to borrow language from the quality improvement literature, there are various 'content theories' for telehealth (Parry et al. 2013).⁷² Also, certain approaches to telehealth may be more effective for some patients than others. As evidence is limited about which of these mechanisms is most likely to be realised and for which patients, ideally evaluations would accumulate evidence to inform future choices about the design of interventions and patient selection (potentially within the formative evaluation framework described below). In the case of WSD, some intermediate outcomes were tested (Steventon et al. 2014), but the relationships between these and the primary endpoint of hospital admission were not assessed. Methods exist from both the causal inference and the social sciences literatures to estimate direct and indirect effects of a treatment in the presence of a single mediator (De Stavola et al. 2015) or a small number of mediators (Daniel et al. 2014). These methods could be applied to test whether the content theory behind telehealth interventions is being realised in practice (Trief et al. 2009).

⁷² Content theory is the rationale for how any improvement in the process measures that is associated with applying the new model of care will lead to improvements in patient outcomes. The other part of a logic model is the 'execution theory', *i.e.*, how the improvement in process measures will be obtained.

A promising strand of research considers the extent to which the medical information collected from telehealth systems can predict events such as hospital admissions (Hamad, Crooks, and Morice 2015). This is important because the logic model of all telemonitoring interventions assumes that the data collected have predictive power. Previously, case-control methods have been used to compare telemonitored data between hospitalised and non-hospitalised patients with heart failure (Chaudhry et al. 2007b), and I am currently collaborating on an application of these methods to a recent RCT (Chaudhry et al. 2010). Research is also needed to examine the role of context in mediating treatment effects (Bate et al. 2014).

8.6.4 Formative evaluation

The case studies in this thesis did not report convincing evidence that the desired reductions in hospital admissions were achieved. This is in line with many other studies in this area (Purdy 2010). One possibility is that complex interventions, with their many elements, require a substantial degree of refinement before they become effective. Therefore, an evaluation model that provides real-time data to enable learning and improvement might have value alongside evaluations that are done as a 'one off' exercise. A potential advantage of working with routinely collected data is that regular refreshes of the data can be made available for research, provided the information infrastructure allows this. Therefore, many of the methods discussed in this thesis could be extended to allow for the ongoing monitoring of effectiveness. If updated estimates were made available to those people delivering the interventions, then the approach could be described as 'theory-driven, formative' evaluation rather than 'summative' evaluation (Scriven 1966).

Some of the issues that need to be addressed in developing formative evaluation approaches relate to the correspondence between the research questions and the decisions that must be made by people developing health care services. For example, the case studies in this thesis addressed questions about 'did it work?' but answering these questions provides limited insight into whether decision makers should, for example, refine the selection of patients, improve implementation to increase fidelity to the original intervention design, alter aspects of the design of the intervention, change the context in which the intervention operates, or stop altogether. Moreover, the degree of belief required for each of these actions might vary depending on whether the intervention is at its early, developmental stage or at its later, scale-up stage (Patton 2011). The multi-phase approach to evaluation described by the Medical Research Council does not fully capture these complexities, for example because it does not assess the role that context can play in determining effectiveness (Craig et al. 2008).

Hargreaves outlined three approaches to rapid evaluation, namely quality improvement, rapid-cycle evaluation, and systems change evaluation (Hargreaves 2014). The Center for Medicare and Medicaid Innovation applied rapid-cycle evaluation methods to its Accountable Care Organisation programme, largely by repeating matched control analyses at regular intervals within the context of a network that

facilitated learning between similar sites (Shrank 2013). This analytical approach does not reflect the multifaceted nature of decision making as described above. Enhancements may be possible by drawing on developments in other fields. For example, the literature on adaptive trial designs provides approaches to inform changes to rates of patient recruitment (Mehta 2013), while some methodological developments for multi-arm trials may be adapted to allow for the multiplicity of health care services offered within a geographic area (Parmar, Carpenter, and Sydes 2014). Decision-theoretic approaches might be helpful to inform decision makers about the trade-offs that they must make (Forster and Pertile 2013), while Bayesian methods (Spiegelhalter et al. 1999) could incorporate prior beliefs from the literature and expert opinion. These methods do not seem to have been widely applied in the evaluation of complex interventions, but could be explored. Regardless of whether these methods are applied within RCTs or observational studies, the focus should be on reducing the gap between the nature and quality of the information provided by the study and the information likely to be most useful to those people who are making decisions about health care services (Patton 2008). It was not an aim of this thesis to develop methods for formative evaluation, but some of the contributions made by this thesis, for example the development of methods to increase the validity of causal estimates from routinely collected data, will facilitate future work in this area.

8.7 Conclusions

By examining complex interventions, this thesis identified particular challenges to evaluation methods that have not been well addressed elsewhere, namely the selection of control areas in observational studies and the assessment of the generalisability of treatments delivered in RCTs. This thesis extended methods for both observational studies and randomised controlled trials, and has applications to evaluations in health services research and beyond.

Based on the case studies examined in this thesis and the wider literature, it seems that many complex interventions are failing to achieve the anticipated reductions in emergency hospital admissions for people with long-term health conditions. This may be because the interventions were targeted at patients unlikely to benefit; the hypothesized mechanisms for effect were incorrect; there were problems with implementation; or the contexts in which the interventions operated were not conducive to the changes being proposed. In order to provide robust evaluations of complex interventions that can help policymakers allocate resources to maximise population health within constrained budgets, further development of evaluation methods is therefore necessary.

References

- Abadie, A., A. Diamond, and J. Hainmueller. 2010. "Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California's Tobacco Control Program." *Journal of the American Statistical Association*, 105(490): 493–505, doi:10.1198/jasa.2009.ap08746.
- Abadie, A. and G. W. Imbens. 2012. "A Martingale Representation for Matching Estimators." *Journal of the American Statistical Association*, 107(498): 833–843, doi:10.1080/01621459.2012.682537.
- Acheson, E. D. and J. G. Evans. 1964. "The Oxford Record Linkage Study: A Review of the Method with Some Preliminary Results." *Proceedings of the Royal Society of Medicine*, 57(4): 269–274.
- Adams, S. T. and S. H. Leveson. 2012. "Clinical Prediction Rules." *BMJ*, 344: d8312, doi:10.1136/bmj.d8312.
- Agresti, A. and Y. Min. 2004. "Effects and Non-Effects of Paired Identical Observations in Comparing Proportions with Binary Matched-Pairs Data." *Statistics in Medicine*, 23(1): 65–75, doi:10.1002/sim.1589.
- Alkema, G. E., K. H. Wilber, G. R. Shannon, and D. Allen. 2007. "Reduced Mortality: The Unexpected Impact of a Telephone-Based Care Management Intervention for Older Adults in Managed Care." *Health Services Research*, 42(4): 1632–1650, doi:10.1111/j.1475-6773.2006.00668.x.
- Austin, P. C. 2008a. "A Critical Appraisal of Propensity-Score Matching in the Medical Literature between 1996 and 2003." *Statistics in Medicine*, 27: 2037–2049, doi:10.1002/sim.3150.
- Austin, P. C. 2009a. "Balance Diagnostics for Comparing the Distribution of Baseline Covariates between Treatment Groups in Propensity-Score Matched Samples." *Statistics in Medicine*, 28: 3083–3107, doi:10.1002/sim.3697.
- Austin, P. C. 2008b. "Discussion of 'A Critical Appraisal of Propensity-Score Matching in the Medical Literature between 1996 and 2003' [rejoinder]." *Statistics in Medicine*, 27(12): 2066–2069, doi:10.1002/sim.3243.
- Austin, P. C. 2010. "Statistical Criteria for Selecting the Optimal Number of Untreated Subjects Matched to Each Treated Subject When Using Many-to-One Matching on the Propensity Score." *American Journal of Epidemiology*, 172(9): 1092–1097, doi:10.1093/aje/kwq224.

- Austin, P. C. 2009b. "The Relative Ability of Different Propensity Score Methods to Balance Measured Covariates between Treated and Untreated Subjects in Observational Studies." *Medical Decision Making*, 29(6): 661–677, doi:10.1177/0272989X09341755.
- Austin, P. C., P. Grootendorst, and G. M. Anderson. 2007. "A Comparison of the Ability of Different Propensity Score Models to Balance Measured Variables between Treated and Untreated Subjects: A Monte Carlo Study." *Statistics in Medicine*, 26(4): 734–753, doi:10.1002/sim.2580.
- Austin, P. C., M. M. Mamdani, T. A. Stukel, G. M. Anderson, and J. V. Tu. 2005. "The Use of the Propensity Score for Estimating Treatment Effects: Administrative versus Clinical Data." *Statistics in Medicine*, 24(10): 1563–1578, doi:10.1002/sim.2053.
- Aylin, P., A. Bottle, and A. Majeed. 2007. "Use of Administrative Data or Clinical Databases as Predictors of Risk of Death in Hospital: Comparison of Models." *BMJ*, 334(7602): 1044, doi:10.1136/bmj.39168.496366.55.
- Azarmina, P. and J. Lewis. 2007. "Patient Satisfaction with a Nurse-Led, Telephone-Based Disease Management Service in Birmingham." *Journal of Telemedicine and Telecare*, 13(5): S3–S4, doi:10.1258/135763307781645194.
- Baker, L. C., S. J. Johnson, D. Macaulay, and H. Birnbaum. 2011. "Integrated Telehealth and Care Management Program for Medicare Beneficiaries with Chronic Disease Linked to Savings." *Health Affairs*, 30(9): 1689–1697, doi:10.1377/hlthaff.2011.0216.
- Bang, H. and J. M. Robins. 2005. "Doubly Robust Estimation in Missing Data and Causal Inference Models." *Biometrics*, 61(4): 962–973, doi:10.1111/j.1541-0420.2005.00377.x.
- Bardsley, M., J. Billings, J. Dixon, T. Georghiou, G. H. Lewis, and A. Steventon. 2011. "Predicting Who Will Use Intensive Social Care: Case Finding Tools Based on Linked Health and Social Care Data." *Age and Ageing*, 40(2): 265–270, doi:10.1093/ageing/afq181.
- Bardsley, M., I. Blunt, S. Davies, and J. Dixon. 2013. "Is Secondary Preventive Care Improving? Observational Study of 10-Year Trends in Emergency Admissions for Conditions Amenable to Ambulatory Care." *BMJ Open*, 3(1): e002007, doi:10.1136/bmjopen-2012-002007.
- Bardsley, M., A. Steventon, and H. Doll. 2013. "Impact of Telehealth on General Practice Contacts: Findings from the Whole Systems Demonstrator Cluster Randomised Trial." *BMC Health Services Research*, 13: 395, doi:10.1186/1472-6963-13-395.

- Barlow, J., S. Bayer, R. Curry, and J. Hendy. 2007a. "The Costs of Telecare: From Pilots to Mainstream Implementation." In *Unit Costs of Health and Social Care 2007*, edited by L. Curtis. Canterbury: Personal Social Services Research Unit.
- Barlow, J., D. Singh, S. Bayer, and R. Curry. 2007b. "A Systematic Review of the Benefits of Home Telecare for Frail Elderly People and Those with Long-Term Conditions." *Journal of Telemedicine and Telecare*, 13(4): 172–179, doi:10.1258/135763307780908058.
- Barnett, T. E., N. R. Chumbler, W. B. Vogel, R. J. Beyth, H. Qin, and R. Kobb. 2006. "The Effectiveness of a Care Coordination Home Telehealth Program for Veterans with Diabetes Mellitus: A 2-Year Follow-Up." *The American Journal of Managed Care*, 12(8): 467–474.
- Baron, J., H. McBain, and S. Newman. 2012. "The Impact of Mobile Monitoring Technologies on Glycosylated Hemoglobin in Diabetes: A Systematic Review." *Journal of Diabetes Science and Technology*, 6(5): 1185–1596, doi:10.1177/193229681200600524.
- Basu, A. 2012. "Patient-Centered or 'Central' Patient: Raising the Veil of Ignorance over Randomization." *Statistics in Medicine*, 31(25): 3057–3059, doi:10.1002/sim.5398.
- Bate, P., G. Robert, N. Fulop, J. Øvretveit, and M. Dixon-Woods. 2014. *Perspectives on Context* edited by J. R. Bamber, London, UK: Health Foundation.
- Begg, C., M. Cho, S. Eastwood, R. Horton, D. Moher, I. Olkin, R. Pitkin, D. Rennie, K. F. Schulz, D. Simel, and D. F. Stroup. 1996. "Improving the Quality of Randomized Controlled Trials: The CONSORT Statement." *JAMA*, 276(8): 637–639, doi:10.1001/jama.1996.03540080059030.
- Belitser, S. V., E. P. Martens, W. R. Pestman, R. H. H. Groenwold, A. de Boer, and O. H. Klungel. 2011. "Measuring Balance and Model Selection in Propensity Score Methods." *Pharmacoepidemiology and Drug Safety*, 20(11): 1115–1129, doi:10.1002/pds.2188.
- Bellón, J. Á., P. Lardelli, J. de D. Luna, and A. Delgado. 2000. "Validity of Self Reported Utilisation of Primary Health Care Services in an Urban Population in Spain." *Journal of Epidemiology and Community Health*, 54(7): 544–551, doi:10.1136/jech.54.7.544.
- Bergmo, T. S. 2009. "Can Economic Evaluation in Telemedicine Be Trusted? A Systematic Review of the Literature." *Cost Effectiveness and Resource Allocation*, 7: 18, doi:10.1186/1478-7547-7-18.
- Berwick, D. M., T. W. Nolan, and J. Whittington. 2008. "The Triple Aim: Care, Health, and Cost." *Health Affairs*, 27(3): 759–769, doi:10.1377/hlthaff.27.3.759.

- Billings, J., J. Dixon, T. Mijanovich, and D. Wennberg. 2006. "Case Finding for Patients at Risk of Readmission to Hospital: Development of Algorithm to Identify High Risk Patients." *BMJ*, 333(7563): 327, doi:10.1136/bmj.38870.657917.AE.
- Billings, J., T. Georghiou, I. Blunt, and M. Bardsley. 2013. "Choosing a Model to Predict Hospital Admission: An Observational Study of New Variants of Predictive Models for Case Finding." *BMJ Open*, 3: e003352, doi:10.1136/bmjopen-2013-003352.
- Birmingham OwnHealth. 2007. *OwnHealth Birmingham: Successes and Learning from the First Year*, Birmingham: Birmingham East and North Primary Care Trust, NHS Direct, Pfizer Health Solutions.
- Black, N., M. Barker, and M. Payne. 2004. "Cross Sectional Survey of Multicentre Clinical Databases in the United." *BMJ*, 328(7454): 1478, doi:10.1136/bmj.328.7454.1478.
- Bourbeau, J., M. Julien, M. François, M. Rouleau, A. Beaupré, R. Bégin, P. Renzi, D. Nault, E. Borycki, K. Schwartzman, R. Singh, and J.-P. Collet. 2003. "Reduction of Hospital Utilization in Patients With Chronic Obstructive Pulmonary Disease." *Archives of Internal Medicine*, 163(5): 585–591, doi:10.1001/archinte.163.5.585.
- Bower, P., M. Cartwright, S. P. Hirani, J. Barlow, J. Hendy, M. Knapp, C. Henderson, A. Rogers, C. Sanders, M. Bardsley, A. Steventon, R. Fitzpatrick, H. Doll, and S. Newman. 2011. "A Comprehensive Evaluation of the Impact of Telemonitoring in Patients with Long-Term Conditions and Social Care Needs: Protocol for the Whole Systems Demonstrator Cluster Randomised Trial." *BMC Health Services Research*, 11: 184, doi:10.1186/1472-6963-11-184.
- Bradford Hill, A. 1937. *Principles of Medical Statistics*, Lancet, London: Lancet.
- Bradley, C. J., L. Penberthy, K. J. Devers, and D. J. Holden. 2010. "Health Services Research and Data Linkages: Issues, Methods, and Directions for the Future." *Health Services Research*, 45(5 Pt 2): 1468–1488, doi:10.1111/j.1475-6773.2010.01142.x.
- Brewin, C. R. and C. Bradley. 1989. "Patient Preferences and Randomised Clinical Trials." *BMJ*, 299(6694): 313–315, doi:10.1136/bmj.299.6694.313.
- Brookhart, M. A., S. Schneeweiss, K. J. Rothman, R. J. Glynn, J. Avorn, and T. Stürmer. 2006. "Variable Selection for Propensity Score Models." *American Journal of Epidemiology*, 163(12): 1149–1156, doi:10.1093/aje/kwj149.
- Burroughs, A. K., C. A. Sabin, K. Rolles, V. Delvart, V. Karam, J. Buckels, J. G. O'Grady, D. Castaing, J. Klempnauer, N. Jamieson, P. Neuhaus, J. Lerut, J. de Ville de Goyet, S. Pollard, M.

- Salizzoni, X. Rogiers, F. Muhlbacher, J. C. G. Valdecasas, C. Broelsch, D. Jaeck, J. Berenguer, E. M. Gonzalez, and R. Adam. 2006. "3-Month and 12-Month Mortality after First Liver Transplant in Adults in Europe: Predictive Models for Outcome." *Lancet*, 367(9506): 225–232, doi:10.1016/S0140-6736(06)68033-1.
- Campbell. 1969. "Perspective: Artifact and Control." In *Artifacts in Behavioral Research*, edited by R. Rosenthal and R. L. Rosnow, pp. 264–288. Oxford, UK: Oxford University Press.
- Car, J., K. Huckvale, and H. Hermens. 2012. "Telehealth for Long Term Conditions." *BMJ*, 344: e4201, doi:10.1136/bmj.e4201.
- Cartwright, M., S. P. Hirani, L. Rixon, M. Beynon, H. Doll, P. Bower, M. Bardsley, A. Steventon, M. Knapp, C. Henderson, A. Rogers, C. Sanders, R. Fitzpatrick, J. Barlow, and S. P. Newman. 2013. "Effect of Telehealth on Quality of Life and Psychological Outcomes over 12 Months (Whole Systems Demonstrator Telehealth Questionnaire Study): Nested Study of Patient Reported Outcomes in a Pragmatic, Cluster Randomised Controlled Trial." *BMJ*, 346: f653, doi:10.1136/bmj.f653.
- Chaudhry, S. I., J. A. Mattera, J. P. Curtis, J. A. Spertus, J. Herrin, Z. Lin, C. O. Phillips, B. V. Hodshon, L. S. Cooper, and H. M. Krumholz. 2010. "Telemonitoring in Patients with Heart Failure." *The New England Journal of Medicine*, 363(24): 2301–2309, doi:10.1056/NEJMoa1010029.
- Chaudhry, S. I., J. A. Mattera, and H. M. Krumholz. 2011. "Telemonitoring in Patients with Heart Failure. Authors' Response to Letters." *New England Journal of Medicine*, 364(11): 1078–1080.
- Chaudhry, S. I., C. O. Phillips, S. S. Stewart, B. J. Riegel, J. A. Mattera, A. F. Jerant, and H. M. Krumholz. 2007a. "Telemonitoring for Patients with Chronic Heart Failure: A Systematic Review." *Journal of Cardiac Failure*, 13(1): 56–62, doi:10.1016/j.cardfail.2006.09.001.
- Chaudhry, S. I., Y. Wang, J. Concato, T. M. Gill, and H. M. Krumholz. 2007b. "Patterns of Weight Change Preceding Hospitalization for Heart Failure." *Circulation*, 116(14): 1549–1554, doi:10.1161/CIRCULATIONAHA.107.690768.
- Chen, H.-F., M. C. Kalish, and J. A. Pagan. 2011. "Telehealth and Hospitalizations for Medicare Home Healthcare Patients." *The American Journal of Managed Care*, 17(6 Spec No.): e224–e230.
- Cherns, A. 1976. "The Principles of Sociotechnical Design." *Human Relations*, 29(8): 783–792.
- Chitnis, X. A., T. Georghiou, A. Steventon, and M. J. Bardsley. 2013. "Effect of a Home-Based End-of-Life Nursing Service on Hospital Use at the End of Life and Place of Death: A Study Using

- Administrative Data and Matched Controls.” *BMJ Supportive & Palliative Care*, 3(4): 422–430, doi:10.1136/bmjspcare-2012-000424.
- Chumbler, N. R., H.-C. Chuang, S. S. Wu, X. Wang, R. Kobb, D. Haggstrom, and H. Jia. 2009. “Mortality Risk for Diabetes Patients in a Care Coordination, Home-Telehealth Programme.” *Journal of Telemedicine and Telecare*, 15(2): 98–101, doi:10.1258/jtt.2008.080803.
- Chumbler, N. R., R. Kobb, D. M. Brennan, and T. Rabinowitz. 2008. “Recommendations for Research Design of Telehealth Studies.” *Telemedicine and e-Health*, 14(9): 986–989, doi:10.1089/tmj.2008.0108.
- Clark, R. A., S. C. Inglis, F. A. McAlister, J. G. F. Cleland, and S. Stewart. 2007. “Telemonitoring or Structured Telephone Support Programmes for Patients with Chronic Heart Failure: Systematic Review and Meta-Analysis.” *BMJ*, 334(7600): 942, doi:10.1136/bmj.39156.536968.55.
- Cochran, W. G. and D. B. Rubin. 1973. “Controlling Bias in Observational Studies: A Review.” *Shankhyā: The Indian Journal of Statistics*, 35(4): 417–446.
- Cochrane, A. 1972. *Effectiveness and Efficiency: Random Reflection on Health Services*, London: Nuffield Provincial Hospitals Trust.
- Cole, S. R. and E. A. Stuart. 2010. “Generalizing Evidence from Randomized Clinical Trials to Target Populations: The ACTG 320 Trial.” *American Journal of Epidemiology*, 172(1): 107–115, doi:10.1093/aje/kwq084.
- Communities and Local Government. 2008. *The English Indices of Deprivation 2007*, London: Communities and Local Government.
- Cook, T. D. and D. T. Campbell. 1979. *Quasi-Experimentation: Design and Analysis Issues for Field Settings*, Boston: Houghton Mifflin.
- Cook, T. D., W. R. Shadish, and V. C. Wong. 2008. “Three Conditions under Which Experiments and Observational Studies Produce Comparable Causal Estimates: New Findings from within-Study Comparisons.” *Journal of Policy Analysis and Management*, 27(4): 724–750, doi:10.1002/pam.20375.
- Cowpertwait, P. S. P. and A. V. Metcalfe. 2009. *Introductory Time Series with R*, Dordrecht: Springer.
- Cox, D. R. and D. Oakes. 1984. *Analysis of Survival Data*, London: Chapman and Hall.

- De Craen, A. J. M., P. J. Roos, A. L. de Vries, and J. Kleijnen. 1996. "Effect of Colour of Drugs: Systematic Review of Perceived Effect of Drugs and of Their Effectiveness." *BMJ*, 313(December): 1624–1626.
- Craig, P., P. Dieppe, S. Macintyre, S. Michie, I. Nazareth, and M. Petticrew. 2008. *Developing and Evaluating Complex Interventions: New Guidance*, London: Medical Research Council.
- Cruickshank, J., G. Beer, E. Winpenny, and J. Manning. 2010. *Healthcare without Walls: A Framework for Delivering Telehealth at Scale*, London: 2020 Health.
- D'Avolio, L., R. Ferguson, S. Goryachev, P. Woods, T. Sabin, J. O'Neil, C. Conrad, J. Gillon, J. Escalera, M. Brophy, P. Lavori, and L. Fiore. 2012. "Implementation of the Department of Veterans Affairs' First Point-of-Care Clinical Trial." *Journal of the American Medical Informatics Association*, 19(e1): e170–6, doi:10.1136/amiajnl-2011-000623.
- Daniel, R. M., B. L. de Stavola, S. N. Cousens, and S. Vansteelandt. 2014. "Causal Mediation Analysis with Multiple Mediators." *Biometrics*, Epub ahead, doi:10.1111/biom.12248.
- Darkins, A., S. Kendall, E. Edmonson, M. Young, and P. Stresel. 2015. "Reduced Cost and Mortality Using Home Telehealth to Promote Self-Management of Complex Chronic Conditions: A Retrospective Matched Cohort Study of 4,999 Veteran Patients." *Telemedicine and e-Health*, 21(1): 70–76, doi:10.1089/tmj.2014.0067.
- Darkins, A., P. Ryan, R. Kobb, L. Foster, E. Edmonson, B. Wakefield, and A. E. Lancaster. 2008. "Care Coordination/Home Telehealth: The Systematic Implementation of Health Informatics, Home Telehealth, and Disease Management to Support the Care of Veteran Patients with Chronic Conditions." *Telemedicine and e-Health*, 14(10): 1118–1126, doi:10.1089/tmj.2008.0021.
- Deeks, J. J., J. Dinnes, R. D'Amico, A. J. Sowden, C. Sakarovitch, F. Song, M. Petticrew, and D. G. Altman. 2003. "Evaluating Non-Randomised Intervention Studies." *Health Technology Assessment*, 7(27): 1–188.
- Dehejia, R. H. and S. Wahba. 2002. "Propensity Score-Matching Methods for Nonexperimental Causal Studies." *Review of Economics and Statistics*, 84(1): 151–161, doi:10.1162/003465302317331982.
- DeMets, D. L. 2012. "Current Development in Clinical Trials: Issues Old and New." *Statistics in Medicine*, 31(25): 2944–2254, doi:10.1002/sim.5405.

- Dendukuri, N., J. McCusker, F. Bellavance, S. Cardin, J. Verdon, I. Karp, and É. Belzile. 2005. "Comparing the Validity of Different Sources of Information on Emergency Department Visits: A Latent Class Analysis." *Medical Care*, 43(3): 266–275.
- Department of Health. 2012. A Concordat between the Department of Health and the Telehealth and Telecare Industry, London: Department of Health.
- Department of Health. 2006. Our Health, Our Care, Our Say: A New Direction for Community Services. Cm 6737, London: The Stationery Office.
- Department of Health. 2007. Payment by Results: Guidance and Tariff for 2008-09, London: Department of Health.
- Department of Health. 2014. Protecting Health and Care Information: A Consultation on Proposals to Introduce New Regulations, Leeds: Department of Health.
- Department of Health. 2011. Whole System Demonstrator Programme: Headline Findings - December 2011, London: Department of Health.
- Diamond, A. and J. S. Sekhon. 2013. "Genetic Matching for Estimating Causal Effects: A General Multivariate Matching Method for Achieving Balance in Observational Studies." *Review of Economics and Statistics*, 95(3): 932–945, doi:10.1162/REST_a_00318.
- Dixon-Woods, M., M. Leslie, C. Tarrant, and J. Bion. 2013. "Explaining Matching Michigan: An Ethnographic Study of a Patient Safety Program." *Implementation Science*, 8: 70, doi:10.1186/1748-5908-8-70.
- Domingo, M., J. Lupón, B. González, E. Crespo, R. López, A. Ramos, A. Urrutia, G. Pera, J. M. Verdú, and A. Bayes-Genis. 2011. "Noninvasive Remote Telemonitoring for Ambulatory Patients With Heart Failure: Effect on Number of Hospitalizations, Days in Hospital, and Quality of Life. CARME (Catalan Remote Management Evaluation) Study." *Revista Española de Cardiología (English Edition)*, 64(4): 277–285, doi:10.1016/j.rec.2010.10.032.
- Drake, C. 1993. "Effects of Misspecification of the Propensity Score on Estimators of Treatment Effect." *Biometrics*, 49(4): 1231–1236.
- Dusheiko, M., H. Gravelle, S. Martin, N. Rice, and P. C. Smith. 2011. "Does Better Disease Management in Primary Care Reduce Hospital Costs? Evidence from English Primary Care." *Journal of Health Economics*, 30(5): 919–932, doi:10.1016/j.jhealeco.2011.08.001.

- Dwinger, S., J. Dirmaier, L. Herbarth, H.-H. König, M. Eckardt, L. Kriston, I. Bermejo, and M. Härter. 2013. "Telephone-Based Health Coaching for Chronically Ill Patients: Study Protocol for a Randomized Controlled Trial." *Trials*, 14: 337, doi:10.1186/1745-6215-14-337.
- Von Elm, E., D. G. Altman, M. Egger, S. J. Pocock, P. C. Gøtzsche, and J. P. Vandenbroucke. 2007. "Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) Statement: Guidelines for Reporting Observational Studies." *BMJ*, 335(7624): 806–808, doi:10.1136/bmj.39335.541782.AD.
- Ettelt, S., N. Mays, and P. Allen. 2014. "The Multiple Purposes of Policy Piloting and Their Consequences: Three Examples from National Health and Social Care Policy in England." *Journal of Social Policy*, 44(2): 319–337, doi:10.1017/S0047279414000865.
- Everett, W., J. C. Kvedar, and T. C. Nesbitt. 2011. "Telemonitoring in Patients with Heart Failure." *New England Journal of Medicine*, 364(11): 1079, doi:10.1056/NEJMc1100395#SA3.
- Fairbrother, P., J. Ure, J. Hanley, L. McCloughan, M. Denvir, A. Sheikh, and B. McKinstry. 2013. "Telemonitoring for Chronic Heart Failure: The Views of Patients and Healthcare Professionals - a Qualitative Study." *Journal of Clinical Nursing*, 23(1-2): 132–144, doi:10.1111/jocn.12137.
- Farmer, A., O. J. Gibson, L. Tarassenko, and A. Neil. 2005. "A Systematic Review of Telemedicine Interventions to Support Blood Glucose Self-Monitoring in Diabetes." *Diabetic Medicine*, 22(10): 1372–1378, doi:10.1111/j.1464-5491.2005.01627.x.
- Farrero, E., J. Escarrabill, E. Prats, M. Maderal, and F. Manresa. 2001. "Impact of a Hospital-Based Home-Care Program on the Management of COPD Patients Receiving Long-Term Oxygen Therapy." *Chest*, 119(2): 364–369, doi:10.1378/chest.119.2.364.
- Fifield, J., D. D. Forrest, M. Martin-Peele, J. A. Burleson, J. Goyzueta, M. Fujimoto, and W. Gillespie. 2013. "A Randomized, Controlled Trial of Implementing the Patient-Centered Medical Home Model in Solo and Small Practices." *Journal of General Internal Medicine*, 28(6): 770–777, doi:10.1007/s11606-012-2197-z.
- Fiore, L. D., M. Brophy, R. E. Ferguson, L. D'Avolio, J. A. Hernos, R. A. Lew, G. Doros, C. H. Conrad, J. A. J. O Neil, T. P. Sabin, J. Kaufman, S. L. Swartz, E. Lawler, M. H. Liang, J. M. Gaziano, and P. W. Lavori. 2011. "A Point-of-Care Clinical Trial Comparing Insulin Administered Using a Sliding Scale versus a Weight-Based Regimen." *Clinical Trials*, 8(2): 183–195, doi:10.1177/1740774510395635.

- Fisher, E. S., D. C. Goodman, and A. Chandra. 2008. *Regional and Racial Variation in Health Care among Medicare Beneficiaries, Lebanon*: The Dartmouth Institute for Health Policy and Clinical Practice.
- Fitzpatrick, R., A. Fletcher, S. Gore, D. Jones, D. Spiegelhalter, and D. Cox. 1992. "Quality of Life Measures in Health Care. I: Applications and Issues in Assessment." *BMJ*, 305(6861): 1074–1077.
- Flather, M., N. Delahunty, and J. Collinson. 2006. "Generalizing Results of Randomized Trials to Clinical Practice: Reliability and Cautions." *Clinical Trials*, 3(6): 508–512, doi:10.1177/1740774506073464.
- Flury, B. and H. Reidwyl. 1986. "Standard Distance in Univariate and Multivariate Analysis." *American Statistician*, 40: 249–51.
- Forster, M. and P. Pertile. 2013. "Optimal Decision Rules for HTA under Uncertainty: A Wider, Dynamic Perspective." *Health Economics*, 22(12): 1507–1514, doi:10.1002/hec.2893.
- Frangakis, C. E. 2009. "The Calibration of Treatment Effects from Clinical Trials to Target Populations." *Clinical Trials*, 6(2): 136–140, doi:10.1177/1740774509103868.
- Frangakis, C. E. and D. B. Rubin. 2002. "Principal Stratification in Causal Inference." *Biometrics*, 58(1): 21–29, doi:10.1111/j.0006-341X.2002.00021.x.
- Franke, R. H. and J. D. Kaul. 1978. "The Hawthorne Experiments: First Statistical Interpretation." *American Sociological Review*, 43(5): 623–643.
- Freund, T., C. Mahler, A. Erler, J. Gensichen, D. Ose, J. Szecsenyi, and F. Peters-Klimm. 2011. "Identification of Patients Likely to Benefit from Care Management Programs." *The American Journal of Managed Care*, 17(5): 345–352.
- Fröbert, O., B. Lagerqvist, G. K. Olivecrona, E. Omerovic, T. Gudnason, M. Maeng, M. Aasa, O. Angerås, F. Calais, M. Danielewicz, D. Erlinge, L. Hellsten, U. Jensen, A. C. Johansson, A. Käregren, J. Nilsson, L. Robertson, L. Sandhall, I. Sjögren, O. Ostlund, J. Harnek, and S. K. James. 2013. "Thrombus Aspiration during ST-Segment Elevation Myocardial Infarction." *The New England Journal of Medicine*, 369(17): 1587–1597, doi:10.1056/NEJMoa1308789.
- Furse, J., M. Clarke, R. Jones, S. Khemka, and G. Findlay. 2008. "Early Experience in Using Telemonitoring for the Management of Chronic Disease in Primary Care." *Journal of Telemedicine and Telecare*, 14(3): 122–4, doi:10.1258/jtt.2008.003005.

- Gheorghe, A., T. E. Roberts, J. C. Ives, B. R. Fletcher, and M. Calvert. 2013. "Centre Selection for Clinical Trials and the Generalisability of Results: A Mixed Methods Study." *PLoS One*, 8(2): e56560, doi:10.1371/journal.pone.0056560.
- Giordano, A., S. Scalvini, E. Zanelli, U. Corrà, G. L. Longobardi, V. A. Ricci, P. Baiardi, and F. Glisenti. 2009. "Multicenter Randomised Trial on Home-Based Telemanagement to Prevent Hospital Readmission of Patients with Chronic Heart Failure." *International Journal of Cardiology*, 131(2): 192–199, doi:10.1016/j.ijcard.2007.10.027.
- Glazerman, S., D. M. Levy, and D. Myers. 2003. "Nonexperimental versus Experimental Estimates of Earnings Impacts." *The Annals of the American Academy of Political and Social Science*, 589(1): 63–93, doi:10.1177/0002716203254879.
- Glidden, D. V. and E. Vittinghoff. 2004. "Modelling Clustered Survival Data from Multicentre Clinical Trials." *Statistics in Medicine*, 23(3): 369–388, doi:10.1002/sim.1599.
- Gloubeman, S. and B. Zimmerman. 2002. *Complicated and Complex Systems: What Would Successful Reform of Medicare Look Like?*, Saskatoon: Commission on the Future of Health Care in Canada.
- Govan, L., O. Wu, A. Briggs, H. M. Colhoun, C. M. Fischbacher, G. P. Leese, J. A. McKnight, S. Philip, N. Sattar, S. H. Wild, and R. S. Lindsay. 2011. "Achieved Levels of HbA1c and Likelihood of Hospital Admission in People with Type 1 Diabetes in the Scottish Population: A Study from the Scottish Diabetes Research Network Epidemiology Group." *Diabetes Care*, 34(9): 1992–1997, doi:10.2337/dc10-2099.
- Gravelle, H., M. Dusheiko, R. Sheaff, P. Sargent, R. Boaden, S. Pickard, S. Parker, and M. Roland. 2007. "Impact of Case Management (Evercare) on Frail Elderly Patients: Controlled before and after Analysis of Quantitative Outcome Data." *BMJ*, 334(7583): 31, doi:10.1136/bmj.39020.413310.55.
- Greenhalgh, T. 2012. "Whole System Demonstrator Trial: Policy, Politics, and Publication Ethics." *BMJ*, 345: e5280; author reply e5305, doi:10.1136/bmj.e5280.
- Greenhalgh, T., R. Procter, J. Wherton, P. Sugarhood, and S. Shaw. 2012. "The Organising Vision for Telehealth and Telecare: Discourse Analysis." *BMJ Open*, 2(4): e001574, doi:10.1136/bmjopen-2012-001574.
- Griswold, M. E. and A. R. Localio. 2010. "Propensity Score Adjustment with Multilevel Data: Setting Your Sites on Decreasing Selection Bias." *Annals of Internal Medicine*, 152(6): 393–396, doi:10.7326/0003-4819-152-6-201003160-00010.

- Groenwold, R. H. H., E. Hak, and A. W. Hoes. 2009. "Quantitative Assessment of Unobserved Confounding Is Mandatory in Nonrandomized Intervention Studies." *Journal of Clinical Epidemiology*, 62(1): 22–28, doi:10.1016/j.jclinepi.2008.02.011.
- Gross, C. P., H. M. Krumholz, G. Van Wye, E. J. Emanuel, and D. Wendler. 2006. "Does Random Treatment Assignment Cause Harm to Research Participants?" *PLoS Medicine*, 3(6): e188, doi:10.1371/journal.pmed.0030188.
- Guo, J., R. T. Konetzka, E. Magett, and W. Dale. 2015. "Quantifying Long-Term Care Preferences." *Medical Decision Making*, 35(1): 106–113, doi:10.1177/0272989X14551641.
- Hamad, G. A., M. Crooks, and A. H. Morice. 2015. "The Value of Telehealth in the Early Detection of Chronic Obstructive Pulmonary Disease Exacerbations: A Prospective Observational Study." *Health Informatics Journal*, Epub Ahead: 1–8, doi:10.1177/1460458214564434.
- Hansen, B. B. 2008a. "The Essential Role of Balance Tests in Propensity-Matched Observational Studies: Comments on 'A Critical Appraisal of Propensity-Score Matching in the Medical Literature between 1996 and 2003' by Peter Austin, *Statistics in Medicine*." *Statistics in Medicine*, 27(12): 2050–2054; discussion 2066–2069, doi:10.1002/sim.3208.
- Hansen, B. B. 2008b. "The Prognostic Analogue of the Propensity Score." *Biometrika*, 95(2): 481–488, doi:10.1093/biomet/asn004.
- Hargreaves, M. B. 2014. *Rapid Evaluation Approaches for Complex Initiatives*, Cambridge, MA: Mathematica.
- Harrell, F. E. J., K. L. Lee, and D. B. Mark. 1996. "Tutorial in Biostatistics. Multivariate Prognostic Models: Issues in Developing Models, Evaluating Assumptions and Adequacy, and Measuring and Reducing Errors." *Statistics in Medicine*, 15(4): 361–387.
- Harrison, D. A., H. Gao, C. A. Welch, and K. M. Rowan. 2010. "The Effects of Critical Care Outreach Services before and after Critical Care: A Matched-Cohort Analysis." *Journal of Critical Care*, 25(2): 196–204, doi:10.1016/j.jcrrc.2009.12.015.
- Hartman, E., R. Grieve, R. Ramsahai, and J. S. Sekhon. 2015. "From Sample Average Treatment Effect to Population Average Treatment Effect on the Treated: Combining Experimental with Observational Studies to Estimate Population Treatment Effects." *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, Epub Ahead, doi:10.1111/rssa.12094.

- Hartman, E., R. Grieve, and J. S. Sekhon. 2010. *From SATE to PATT: The Essential Role of Placebo Tests for Combining Experimental with Observational Studies*, Berkeley: University of California, Berkeley.
- Hartman, E. and F. D. Hidalgo. 2011. *What's the Alternative? An Equivalence Approach to Balance and Placebo Tests*, Berkeley: University of California, Berkeley.
- Hauck, K. and X. Zhao. 2011. "How Dangerous Is a Day in Hospital? A Model of Adverse Events and Length of Stay for Medical Inpatients." *Medical Care*, 49(12): 1068–1075, doi:10.1097/MLR.0b013e31822efb09.
- Hayes, R. J. and S. Bennett. 1999. "Simple Sample Size Calculation for Cluster-Randomized Trials." *International Journal of Epidemiology*, 28(2): 319–326, doi:10.1093/ije/28.2.319.
- Haynes, B. 1999. "Can It Work? Does It Work? Is It Worth It?" *BMJ*, 319(7211): 652–653.
- Health & Social Care Information Centre. 2013. "Secondary Uses Service (SUS)," Available at: <http://www.hscic.gov.uk/sus>.
- Heckman, J. J. 1979. "Sample Selection Bias as a Specification Error." *Econometrica*, 47(1): 153–161, doi:10.2307/1912352.
- Heckman, J. J. and J. A. Smith. 1999. "The Pre- programme Earnings Dip and the Determinants of Participation in a Social Programme. Implications for Simple Programme Evaluation Strategies." *The Economic Journal*, 109(457): 313–348.
- Heckman, J. J., S. Urzua, and E. Vytlacil. 2006. "Understanding Instrumental Variables in Models with Essential Heterogeneity." *The Review of Economics and Statistics*, 88(3): 387–403.
- Henderson, C., M. Knapp, J.-L. Fernández, J. Beecham, S. P. Hirani, M. Beynon, M. Cartwright, L. Rixon, H. Doll, P. Bower, A. Steventon, A. Rogers, R. Fitzpatrick, J. Barlow, M. Bardsley, and S. P. Newman. 2014. "Cost-Effectiveness of Telecare for People with Social Care Needs: The Whole Systems Demonstrator Cluster Randomised Trial." *Age and Ageing*, 43(6): 794–800, doi:10.1093/ageing/afu067.
- Henderson, C., M. Knapp, J.-L. Fernandez, J. Beecham, S. P. Hirani, M. Cartwright, L. Rixon, M. Beynon, A. Rogers, P. Bower, H. Doll, R. Fitzpatrick, A. Steventon, M. Bardsley, J. Hendy, and S. P. Newman. 2013. "Cost Effectiveness of Telehealth for Patients with Long Term Conditions (Whole Systems Demonstrator Telehealth Questionnaire Study): Nested Economic Evaluation in a Pragmatic, Cluster Randomised Controlled Trial." *BMJ*, 346: f1035, doi:10.1136/bmj.f1035.

- Hendy, J., T. Chrysanthaki, J. Barlow, M. Knapp, A. Rogers, C. Sanders, P. Bower, R. Bowen, R. Fitzpatrick, M. Bardsley, and S. Newman. 2012. "An Organisational Analysis of the Implementation of Telecare and Telehealth: The Whole Systems Demonstrator." *BMC Health Services Research*, 12: 403, doi:10.1186/1472-6963-12-403.
- Hibbard, J. H. and J. Greene. 2013. "What the Evidence Shows about Patient Activation: Better Health Outcomes and Care Experiences; Fewer Data on Costs." *Health Affairs*, 32(2): 207–214, doi:10.1377/hlthaff.2012.1061.
- Higashi, T., P. G. Shekelle, J. L. Adams, C. J. Kamberg, C. P. Roth, D. H. Solomon, D. B. Reuben, L. Chiang, C. H. MacLean, J. T. Chang, R. T. Young, D. M. Saliba, and N. S. Wenger. 2005. "Quality of Care Is Associated with Survival in Vulnerable Older Patients." *Annals of Internal Medicine*, 143(4): 274–281.
- Hill, B., B. Richardson, and H. Skouteris. 2014. "Do We Know How to Design Effective Health Coaching Interventions: A Systematic Review of the State of the Literature." *American Journal of Health Promotion*, Epub Ahead, doi:10.4278/ajhp.130510-LIT-238.
- Hill, J. 2008. "Discussion of Research Using Propensity- score Matching: Comments on 'A Critical Appraisal of Propensity- score Matching in the Medical Literature between 1996 and 2003' by Peter Austin, *Statistics in Medicine*." *Statistics in Medicine*, 27(12): 2055–2061, doi:10.1002/sim.3245.
- Ho, D. E., K. Imai, G. King, and E. A. Stuart. 2007. "Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference." *Political Analysis*, 15(3): 199–236, doi:10.1093/pan/15.3.199.
- Hobbs, S. 2009. "Newham WSD TeleHealth: Improving Quality of Life at Home," NHS Newham, Available at: [http://www.newhamwsdtrial.org/user_files/downloads/PR/PR July - TH Terry Munro.pdf](http://www.newhamwsdtrial.org/user_files/downloads/PR/PR%20July%20-%20TH%20Terry%20Munro.pdf).
- Holland, P. W. 1986. "Statistics and Causal Inference." *Journal of the American Statistical Association*, 81(396): 945–960.
- Hopkinson, N. S. 2012. "Re: Effect of Telehealth on Use of Secondary Care and Mortality: Findings from the Whole System Demonstrator Cluster Randomised Trial." *BMJ*, 344: e3874.
- Hripcsak, G. and D. J. Albers. 2013. "Next-Generation Phenotyping of Electronic Health Records." *Journal of the American Medical Informatics Association*, 20(1): 117–121, doi:10.1136/amiainl-2012-001145.

- Hutchison, A. J. and J. D. Breckon. 2011. "A Review of Telephone Coaching Services for People with Long-Term Conditions." *Journal of Telemedicine and Telecare*, 17(8): 451–458, doi:10.1258/jtt.2011.110513.
- Iezzoni, L. I. 1997. "Assessing Quality Using Administrative Data." *Annals of Internal Medicine*, 127(8 Pt 2): 666–674.
- Imai, K., G. King, and E. A. Stuart. 2008. "Misunderstandings between Experimentalists and Observationalists about Causal Inference." *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 171(2): 481–502, doi:10.1111/j.1467-985X.2007.00527.x.
- Imbens, G. W. and T. Lemieux. 2008. "Regression Discontinuity Designs: A Guide to Practice." *Journal of Econometrics*, 142(2): 615–635, doi:10.1016/j.jeconom.2007.05.001.
- Inglis, S. C., R. A. Clark, F. A. McAlister, J. Ball, C. Lewinter, D. Cullington, S. Stewart, and J. G. F. Cleland. 2010. "Structured Telephone Support or Telemonitoring Programmes for Patients with Chronic Heart Failure." *Cochrane Database of Systematic Reviews*, 8: Art. No. CD007228, doi:10.1002/14651858.CD007228.pub2.
- Jia, H., H.-C. Chuang, S. S. Wu, X. Wang, and N. R. Chumbler. 2009. "Long-Term Effect of Home Telehealth Services on Preventable Hospitalization Use." *Journal of Rehabilitation Research and Development*, 46(5): 557–566.
- Jia, H., H. Feng, X. Wang, S. S. Wu, and N. Chumbler. 2011. "A Longitudinal Study of Health Service Utilization for Diabetes Patients in a Care Coordination Home-Telehealth Programme." *Journal of Telemedicine and Telecare*, 17(3): 123–126, doi:10.1258/jtt.2010.100314.
- Jones, A. M. 2007. "Identification of Treatment Effects in Health Economics." *Health Economics*, 16(11): 1127–1131, doi:10.1002/hec.1302.
- Jonk, Y., K. Lawson, H. O'Connor, K. S. Riise, D. Eisenberg, B. Dowd, and M. J. Kreitzer. 2015. "How Effective Is Health Coaching in Reducing Health Services Expenditures?" *Medical Care*, 53(2): 133–140, doi:10.1097/MLR.0000000000000287.
- Jordan, R. E., R. J. Lancashire, and P. Adab. 2011. "An Evaluation of Birmingham Own Health Telephone Care Management Service among Patients with Poorly Controlled Diabetes. A Retrospective Comparison with the General Practice Research Database." *BMC Public Health*, 11: 707, doi:10.1186/1471-2458-11-707.
- Joynt, K. E. and A. K. Jha. 2012. "Thirty-Day Readmissions—Truth and Consequences." *The New England Journal of Medicine*, 366(15): 1366–1369, doi:10.1056/NEJMp1201598.

- Kaplan, E. L. and P. Meier. 1958. "Nonparametric Estimation from Incomplete Observations." *Journal of the American Statistical Association*, 53(282): 457–481.
- Keele, L. 2010. *An Overview of Rbounds: An R Package for Rosenbaum Bounds Sensitivity Analysis with Matched Data.*, State College: Penn State University.
- Kim, S. E., C. Michalopoulos, R. M. Kwong, A. Warren, and M. S. Manno. 2013. "Telephone Care Management's Effectiveness in Coordinating Care for Medicaid Beneficiaries in Managed Care: A Randomized Controlled Study." *Health Services Research*, 48(5): 1730–1749, doi:10.1111/1475-6773.12060.
- King, G. and L. Zeng. 2006. "The Dangers of Extreme Counterfactuals." *Political Analysis*, 14(2): 131–159, doi:10.1093/pan/mpj004.
- Klug, C., K. Bonin, N. Bultemeier, Y. Rozenfeld, R. S. Vasquez, M. Johnson, and J. C. Cherry. 2011. "Integrating Telehealth Technology into a Clinical Pharmacy Telephonic Diabetes Management Program." *Journal of Diabetes Science and Technology*, 5(5): 1238–1245.
- Kolata, G. 2014. "Method of Study Is Criticized in Group's Health Policy Tests." *New York Times*.
- Kreif, N., R. Grieve, R. Radice, Z. Sadique, R. Ramsahai, and J. S. Sekhon. 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., R. Grieve, R. Radice, and J. S. Sekhon. 2013. "Regression-Adjusted Matching and Double-Robust Methods for Estimating Average Treatment Effects in Health Economic Evaluation." *Health Services and Outcomes Research Methodology*, 13(2-4): 174–202, doi:10.1007/s10742-013-0109-2.
- Kreif, N., S. Gruber, R. Radice, R. Grieve, and J. S. Sekhon. 2014. "Evaluating Treatment Effectiveness under Model Misspecification: A Comparison of Targeted Maximum Likelihood Estimation with Bias-Corrected Matching." *Statistical Methods in Medical Research*, Epub Ahead: 1–22, doi:10.1177/0962280214521341.
- L&M Policy Research LLC. 2013. *Evaluation of CMMI Accountable Care Organization Initiatives*, Washington DC: L&M Policy Research, LLC.
- Lagerqvist, B., O. Fröbert, G. K. Olivecrona, T. Gudnason, M. Maeng, P. Alström, J. Andersson, F. Calais, J. Carlsson, O. Collste, M. Göteborg, P. Hårdhammar, D. Ioanes, A. Kallryd, R. Linder, A. Lundin, J. Odenstedt, E. Omerovic, V. Puskar, T. Tödt, E. Zelleröth, O. Östlund, and S. K.

- James. 2014. "Outcomes 1 Year after Thrombus Aspiration for Myocardial Infarction." *New England Journal of Medicine*, 371(12): 1111–1120, doi:10.1056/NEJMoa1405707.
- LaLonde, R. 1986. "Evaluating the Econometric Evaluations of Training Programs with Experimental Data." *The American Economic Review*, 76(4): 604–620.
- Lancaster, G. A., H. Chellaswamy, S. Taylor, D. Lyon, and C. Dowrick. 2007. "Design of a Clustered Observational Study to Predict Emergency Admissions in the Elderly: Statistical Reasoning in Clinical Practice." *Journal of Evaluation in Clinical Practice*, 13(2): 169–178, doi:10.1111/j.1365-2753.2006.00663.x.
- Larsen, K. and J. Merlo. 2005. "Appropriate Assessment of Neighborhood Effects on Individual Health: Integrating Random and Fixed Effects in Multilevel Logistic Regression." *American Journal of Epidemiology*, 161(1): 81–88, doi:10.1093/aje/kwi017.
- Lauer, M. S. and R. B. J. D'Agostino. 2013. "The Randomized Registry Trial - The Next Disruptive Technology in Clinical Research?" *The New England Journal of Medicine*, 369(17): 1579–1581, doi:10.1056/NEJMp1310102.
- Lee, B. K., J. Lessler, and E. A. Stuart. 2010. "Improving Propensity Score Weighting Using Machine Learning." *Statistics in Medicine*, 29(3): 337–346, doi:10.1002/sim.3782.
- Lee, D., P. Austin, and J. Rouleau. 2003. "Predicting Mortality Among Patients Hospitalized for Heart Failure: Derivation and Validation of a Clinical Model." *JAMA*, 290(19): 2581–2587.
- Lewis, G. H. 2006. Case Study: Virtual Wards at Croydon Primary Care Trust, London: King's Fund.
- Lewis, G. H., T. Georghiou, A. Steventon, R. Vaithianathan, X. Chitnis, J. Billings, I. Blunt, L. Wright, A. Roberts, and M. Bardsley. 2013. Impact of "Virtual Wards" on Hospital Use: A Research Study Using Propensity Matched Controls and a Cost Analysis, Southampton: NIHR Service Delivery and Organisation Programme.
- Lewsey, J. D., A. H. Leyland, G. D. Murray, and F. A. Boddy. 2000. "Using Routine Data to Complement and Enhance the Results of Randomised Controlled Trials." *Health Technology Assessment*, 4(22): 1–55.
- Lin, W.-C., H.-L. Chien, G. Willis, E. O'Connell, K. S. Rennie, H. M. Bottella, and T. G. Ferris. 2012. "The Effect of a Telephone-Based Health Coaching Disease Management Program on Medicaid Members with Chronic Conditions." *Medical Care*, 50(1): 91–98, doi:10.1097/MLR.0b013e31822dcedf.

- Little, R. J. and D. B. Rubin. 2000. "Causal Effects in Clinical and Epidemiological Studies via Potential Outcomes: Concepts and Analytical Approaches." *Annual Review of Public Health*, 21: 121–45, doi:10.1146/annurev.publhealth.21.1.121.
- Lu, B. and P. R. Rosenbaum. 2004. "Optimal Pair Matching with Two Control Groups." *Journal of Computational and Graphical Statistics*, 13(2): 422–434, doi:10.1198/1061860043470.
- Lucas, F. L., A. E. Siewers, D. J. Malenka, and D. E. Wennberg. 2008. "Diagnostic-Therapeutic Cascade Revisited: Coronary Angiography, Coronary Artery Bypass Graft Surgery, and Percutaneous Coronary Intervention in the Modern Era." *Circulation*, 118(25): 2797–802, doi:10.1161/CIRCULATIONAHA.108.789446.
- Mahalanobis, P. C. 1936. "On the Generalised Distance in Statistics." *Proceedings of the National Institute of Sciences of India*, 2(1): 49 – 55.
- Malanda, U. L., L. M. C. Welschen, I. I. Riphagen, J. M. Dekker, G. Nijpels, and S. D. M. Bot. 2012. "Self-Monitoring of Blood Glucose in Patients with Type 2 Diabetes Mellitus Who Are Not Using Insulin." *Cochrane Database of Systematic Reviews*, (2): CD005060, doi:10.1002/14651858.CD005060.pub2.
- Marcus, S. M., E. A. Stuart, P. Wang, W. R. Shadish, and P. M. Steiner. 2012. "Estimating the Causal Effect of Randomization versus Treatment Preference in a a Doubly Randomized Preference Trial." *Psychological Methods*, 17(2): 244–254, doi:10.1037/a0028031.Estimating.
- Mamot, M. and R. G. Wilkinson eds. 2009. *Social Determinants of Health* 2nd ed., Oxford: Oxford University Press.
- Mason, J. M., R. J. Young, J. P. New, J. M. Gibson, A. F. Long, T. Gambling, and T. Friede. 2006. "Economic Analysis of a Telemedicine Intervention to Improve Glycemic Control in Patients with Diabetes Mellitus." *Disease Management & Health Outcomes*, 14(6): 377–385, doi:10.2165/00115677-200614060-00007.
- Mauri, L. 2012. "Why We Still Need Randomized Trials to Compare Effectiveness." *The New England Journal of Medicine*, 366(16): 1538–1540, doi:10.1056/NEJMe1202866.
- McCambridge, J., K. Kypri, and D. Elbourne. 2014. "In Randomization We Trust? There Are Overlooked Problems in Experimenting with People in Behavioral Intervention Trials." *Journal of Clinical Epidemiology*, 67(3): 247–253, doi:10.1016/j.jclinepi.2013.09.004.

- McCambridge, J., A. Sorhaindo, A. Quirk, and K. Nanchahal. 2014. "Patient Preferences and Performance Bias in a Weight Loss Trial with a Usual Care Arm." *Patient Education and Counseling*, 95(2): 243–247, doi:10.1016/j.pec.2014.01.003.
- McCarney, R., J. Warner, S. Iliffe, R. van Haselen, M. Griffin, and P. Fisher. 2007. "The Hawthorne Effect: A Randomised, Controlled Trial." *BMC Medical Research Methodology*, 7: 30, doi:10.1186/1471-2288-7-30.
- McConnell, K. . J., N. T. Wallace, C. A. Gallia, and J. A. Smith. 2008. "Effect of Eliminating Behavioral Health Benefits for Selected Medicaid Enrollees." *Health Services Research*, 43(4): 1348–1365, doi:10.1111/j.1475-6773.2008.00844.x.
- McKinstry, B., J. Hanley, S. Wild, C. Pagliari, M. Paterson, S. Lewis, A. Sheikh, A. Krishan, A. Stoddart, and P. Padfield. 2013. "Telemonitoring Based Service Redesign for the Management of Uncontrolled Hypertension: Multicentre Randomised Controlled Trial." *BMJ*, 346: f3030, doi:10.1136/bmj.f3030.
- McLean, S., D. Chandler, U. Numatov, J. Liu, C. Pagliari, J. Car, and A. Sheikh. 2010. "Telehealthcare for Asthma." *Cochrane Database of Systematic Reviews*, 6(10): Art. No.:CD007717, doi:10.1002/14651858.CD007717.pub2.
- McLean, S., U. Numatov, J. L. Y. Liu, C. Pagliari, J. Car, and A. Sheikh. 2011. "Telehealthcare for Chronic Obstructive Pulmonary Disease (Review)." *Cochrane Database of Systematic Reviews*, (7): Art. No. CD007718, doi:10.1002/14651858.CD007718.pub2.
- McLean, S., D. Protti, and A. Sheikh. 2011. "Telehealthcare for Long Term Conditions." *BMJ*, 342: d120, doi:10.1136/bmj.d120.
- McLean, S., A. Sheikh, K. Cresswell, U. Nurmatov, M. Mukherjee, A. Hemmi, and C. Pagliari. 2013. "The Impact of Telehealthcare on the Quality and Safety of Care: A Systematic Overview." *PLoS One*, 8(8): e71238, doi:10.1371/journal.pone.0071238.
- Mehta, C. R. 2013. "Adaptive Clinical Trial Designs with Pre-Specified Rules for Modifying the Sample Size: A Different Perspective." *Statistics in Medicine*, 32(8): 1276–1279, doi:10.1002/sim.5720.
- Miettinen, O. S. 1985. "The 'Case-Control' Study: Valid Selection of Subjects." *Journal of Chronic Disease*, 38(7): 543–548.

- Ming, K. and P. R. Rosenbaum. 2000. "Substantial Gains in Bias Reduction from Matching with a Variable Number of Controls." *Biometrics*, 56(1): 118–124, doi:10.1111/j.0006-341X.2000.00118.x.
- Morguet, A. J., P. Kühnelt, A. Kallel, M. Jaster, and H.-P. Schultheiss. 2008. "Impact of Telemedical Care and Monitoring on Morbidity in Mild to Moderate Chronic Heart Failure." *Cardiology*, 111(2): 134–139, doi:10.1159/000119701.
- Mosis, G., J. P. Dieleman, B. C. Stricker, J. van der Lei, and M. C. J. M. Sturkenboom. 2006. "A Randomized Database Study in General Practice Yielded Quality Data but Patient Recruitment in Routine Consultation Was Not Practical." *Journal of Clinical Epidemiology*, 59(5): 497–502, doi:10.1016/j.jclinepi.2005.11.007.
- Motheral, B. R. 2011. "Telephone-Based Disease Management: Why It Does Not Save Money." *The American Journal of Managed Care*, 17(1): e10–16.
- Mowles, C. 2014. "Complex, but Not Quite Complex Enough: The Turn to the Complexity Sciences in Evaluation Scholarship." *Evaluation*, 20(2): 160–175, doi:10.1177/1356389014527885.
- Nelson, E. C., P. B. Batalden, and M. M. Godfrey. 2007. *Quality By Design: A Clinical Microsystems Approach* 1st Ed., San Francisco: Jossey-Bass.
- Nelson, E. C., E. Eftimovska, C. Lind, A. Hager, J. H. Wasson, and S. Lindblad. 2015. "Patient Reported Outcome Measures in Practice." *BMJ*, 350: g7818, doi:10.1136/bmj.a2597.
- Nelson, L. 2012. *Lessons from Medicare's Demonstration Projects on Disease Management and Care Coordination*, 2012-01 Working Paper Series, Washington DC: Congressional Budget Office.
- Newhouse, J. P. and M. McClellan. 1998. "Econometrics in Outcomes Research: The Use of Instrumental Variables." *Annual Review of Public Health*, 19: 17–34, doi:10.1146/annurev.publhealth.19.1.17.
- Newman, S. P., M. Bardsley, J. Barlow, J. Beecham, M. Beynon, J. Billings, A. Bowen, P. Bower, M. Cartwright, T. Chrysanthaki, J. Dixon, H. Doll, J.-L. Fernandez, R. Fitzpatrick, C. Henderson, J. Hendy, S. P. Hirani, M. Knapp, L. Rixon, A. Rogers, C. Sanders, L. A. Silva, and A. Steventon. 2014. *The Whole System Demonstrator Programme: Final Report*, London: City University London.
- NHS Commissioning Board. 2013. *Commissioning for Quality and Innovation (CQUIN): 2013/4 Guidance*, Leeds: NHS Commissioning Board.

- NHS Direct. 2011. Chief Executive's Report: Agenda Item 6. 19 December 2011, Leeds: NHS Direct.
- NHS England. 2014. Five Year Forward View, London: NHS England.
- Nilsson, M., U. Rasmak, H. Nordgren, P. Hallberg, J. Skönevik, G. Westman, and O. Rolandsson. 2009. "The Physician at a Distance: The Use of Videoconferencing in the Treatment of Patients with Hypertension." *Journal of Telemedicine and Telecare*, 15(8): 397–403, doi:10.1258/jtt.2009.090509.
- Nolte, E. and S. Hinrichs eds. 2012. DISMEVAL. Developing and Validating Disease Management Evaluation Methods for European Healthcare Systems: Final Report, Santa Monica, CA: RAND Corporation.
- Normand, S.-L. T., M. B. Landrum, E. Guadagnoli, J. Z. Ayanian, T. J. Ryan, P. D. Cleary, and B. J. McNeil. 2001. "Validating Recommendations for Coronary Angiography Following Acute Myocardial Infarction in the Elderly: A Matched Analysis Using Propensity Scores." *Journal of Clinical Epidemiology*, 54(4): 387–398.
- Nuffield Trust. 2011. Predictive Risk and Health Care: An Overview, London: Nuffield Trust.
- Nymark, L. S., P. Davies, O. Shabestari, and I. McNeil. 2013. "Analysis of the Impact of the Birmingham OwnHealth Program on Secondary Care Utilization and Cost: A Retrospective Cohort Study." *Telemedicine and e-Health*, 19(12): 949–955, doi:10.1089/tmj.2013.0011.
- Office for National Statistics. 2010. 2001 Area Classification of Health Areas – Corresponding Areas, London: Office for National Statistics.
- Olschewski, M. and H. Scheurlen. 1985. "Comprehensive Cohort Study: An Alternative to Randomized Consent Design in a Breast Preservation Trial." *Methods of Information in Medicine*, 24(3): 131–4.
- Olschewski, M., M. Schumacher, and K. B. Davis. 1992. "Analysis of Randomized and Nonrandomized Patients in Clinical Trials Using the Comprehensive Cohort Follow-up Study Design." *Controlled Clinical Trials*, 13(3): 226–239, doi:10.1016/0197-2456(92)90005-K.
- Olsen, R. B., L. L. Orr, S. H. Bell, and E. A. Stuart. 2013. "External Validity in Policy Evaluations That Choose Sites Purposively." *Journal of Policy Analysis and Management*, 32(1): 107–121, doi:10.1002/pam.

- Padkin, A., K. Rowan, and N. Black. 2001. "Using High Quality Clinical Databases to Complement the Results of Randomised Controlled Trials: The Case of Recombinant Human Activated Protein C." *BMJ*, 323(7318): 923–926.
- Paré, G., M. Jaana, and C. Sicotte. 2007. "Systematic Review of Home Telemonitoring for Chronic Diseases: The Evidence Base." *Journal of the American Medical Informatics Association*, 14(3): 269–277, doi:http://dx.doi.org/10.1197/jamia.M2270.
- Parmar, M. K. B., J. Carpenter, and M. R. Sydes. 2014. "More Multiarm Randomised Trials of Superiority Are Needed." *Lancet*, 384(9940): 283–284, doi:10.1016/S0140-6736(14)61122-3.
- Parry, G. J., A. Carson-Stevens, D. F. Luff, M. E. McPherson, and D. A. Goldmann. 2013. "Recommendations for Evaluation of Health Care Improvement Initiatives." *Academic Pediatrics*, 13(6 Suppl): S23–S30, doi:10.1016/j.acap.2013.04.007.
- Patton, M. Q. 2011. *Developmental Evaluation: Applying Complexity Concepts to Enhance Innovation and Use* 1st Ed., New York City: The Guilford Press.
- Patton, M. Q. 2008. *Utilization-Focused Evaluation* 4th Ed., Thousand Oaks: SAGE Publications, Inc.
- Pekmezaris, R., I. Mitzner, K. R. Pecinka, C. N. Nouryan, M. L. Lesser, M. Siegel, J. W. Swiderski, G. Moise, R. Younker Sr., and K. Smolich. 2012. "The Impact of Remote Patient Monitoring (telehealth) upon Medicare Beneficiaries with Heart Failure." *Telemedicine and e-Health*, 18(2): 101–108, doi:10.1089/tmj.2011.0095.
- Pinnock, H., J. Hanley, L. McCloughan, A. Todd, A. Krishan, S. Lewis, A. Stoddart, M. van der Pol, W. MacNee, A. Sheikh, C. Pagliari, and B. McKinstry. 2013. "Effectiveness of Telemonitoring Integrated into Existing Clinical Services on Hospital Admission for Exacerbation of Chronic Obstructive Pulmonary Disease: Researcher Blind, Multicentre, Randomised Controlled Trial." *BMJ*, 347: f6070, doi:10.1136/bmj.f6070.
- Polisena, J., K. Tran, K. Cimon, B. Hutton, S. McGill, and K. Palmer. 2009. "Home Telehealth for Diabetes Management: A Systematic Review and Meta-Analysis." *Diabetes, Obesity & Metabolism*, 11(10): 913–930, doi:10.1111/j.1463-1326.2009.01057.x.
- Polisena, J., K. Tran, K. Cimon, B. Hutton, S. McGill, K. Palmer, and R. E. Scott. 2010a. "Home Telehealth for Chronic Obstructive Pulmonary Disease: A Systematic Review and Meta-Analysis." *Journal of telemedicine and Telecare*, 16(3): 120–7, doi:10.1258/jtt.2009.090812.

- Polisena, J., K. Tran, K. Cimon, B. Hutton, S. McGill, K. Palmer, and R. E. Scott. 2010b. "Home Telemonitoring for Congestive Heart Failure: A Systematic Review and Meta-Analysis." *Journal of Telemedicine and Telecare*, 16(2): 68–76, doi:10.1258/jtt.2009.090406.
- Pope, C., S. Halford, R. Tinati, and M. Weal. 2014. "What's the Big Fuss about 'Big Data'?" *Journal of Health Services Research & Policy*, 19(2): 67–68, doi:10.1177/1355819614521181.
- Pressler, T. R. and E. E. Kaizar. 2013. "The Use of Propensity Scores and Observational Data to Estimate Randomized Controlled Trial Generalizability Bias." *Statistics in Medicine*, 32(20): 3552–3568, doi:10.1002/sim.5802.
- Puffer, S., D. J. Torgerson, and J. Watson. 2003. "Evidence for Risk of Bias in Cluster Randomised Trials: Review of Recent Trials Published in Three General Medical Journals." *BMJ*, 327: 785, doi:10.1136/bmj.327.7418.785.
- Purdy, S. 2010. *Avoiding Hospital Admissions: What Does the Research Evidence Say?*, London: The King's Fund.
- Quintana, J. M., C. Esteban, A. Unzurrunzaga, S. Garcia-Gutierrez, N. Gonzalez, I. Barrio, I. Arostegui, I. Lafuente, M. Bare, N. Fernandez-de-Larrea, and S. Vidal. 2014. "Predictive Score for Mortality in Patients with COPD Exacerbations Attending Hospital Emergency Departments." *BMC Medicine*, 12: 66, doi:10.1186/1741-7015-12-66.
- Radice, R., R. Ramsahai, R. Grieve, N. Kreif, Z. Sadique, and J. S. Sekhon. 2012. "Evaluating Treatment Effectiveness in Patient Subgroups: A Comparison of Propensity Score Methods with an Automated Matching Approach." *The International Journal of Biostatistics*, 8(1): 25, doi:10.1515/1557-4679.1382.
- Ramsahai, R. R., R. Grieve, and J. S. Sekhon. 2011. "Extending Iterative Matching Methods: An Approach to Improving Covariate Balance That Allows Prioritisation." *Health Services and Outcomes Research Methodology*, 11(3-4): 95–114, doi:10.1007/s10742-011-0075-5.
- Richards, S. H., J. Coast, and T. J. Peters. 2003. "Patient-Reported Use of Health Service Resources Compared with Information from Health Providers." *Health & Social Care in the Community*, 11(6): 510–518, doi:10.1046/j.1365-2524.2003.00457.x.
- Roberts, C. and D. J. Torgerson. 1999. "Baseline Imbalance in Randomised Controlled Trials." *BMJ*, 319: 185, doi:10.1136/bmj.319.7203.185.

- Robinson, S. 2013. "NHS England Abandons Health Secretary's Pledge on Telehealth," GP Online, Available at: http://www.gponline.com/bulletin/daily_news/article/1222218/nhs-england-abandons-health-secretarys-pledge-telehealth/?DCMP=EMC-ED-News,jobsandCPD-1222218.
- Roland, M., M. Dusheiko, H. Gravelle, and S. Parker. 2005. "Follow up of People Aged 65 and over with a History of Emergency Admissions: Analysis of Routine Admission Data." *BMJ*, 330: 289–292, doi:10.1136/bmj.330.7486.289.
- Roland, M., R. Lewis, A. Steventon, G. Abel, J. Adams, M. Bardsley, L. Brereton, X. Chitnis, A. Conklin, L. Staetsky, S. Tunkel, and T. Ling. 2012. "Case Management for at-Risk Elderly Patients in the English Integrated Care Pilots: Observational Study of Staff and Patient Experience and Secondary Care Utilisation." *International Journal of Integrated Care*, 12: e130.
- Roland, M. and D. J. Torgerson. 1998. "Understanding Controlled Trials: What Are Pragmatic Trials?" *BMJ*, 316(7127): 285, doi:10.1136/bmj.316.7127.285.
- Roos, L. L., C. A. Mustard, J. P. Nicol, D. F. McLerran, D. J. Malenka, T. K. Young, and M. M. Cohen. 1993. "Registries and Administrative Data: Organization and Accuracy." *Medical Care*, 31(3): 201–212.
- Rosenbaum, P. R. 2002. *Observational Studies* 2nd Ed., New York: Springer.
- Rosenbaum, P. R. 1987. "The Role of a Second Control Group in an Observational Study." *Statistical Science*, 2(3): 292–316.
- Rosenbaum, P. R. and D. B. Rubin. 1985. "Constructing a Control Group Using Multivariate Matched Sampling Methods That Incorporate the Propensity Score." *The American Statistician*, 39(1): 33–38.
- Rosenbaum, P. R. and D. B. Rubin. 1983. "The Central Role of the Propensity Score in Observational Studies for Causal Effects." *Biometrika*, 70(1): 41–55.
- Rosnow, R. L. and R. Rosenthal. 1997. *People Studying People: Artifacts and Ethics in Behavioral Research*, New York: WH Freeman.
- Ross, J. S., G. K. Mulvey, B. Stauffer, V. Patlolla, S. M. Bernheim, P. S. Keenan, and H. M. Krumholz. 2008. "Statistical Models and Patient Predictors of Readmission for Heart Failure: A Systematic Review." *Archives of Internal Medicine*, 168(13): 1371–1386, doi:10.1001/archinte.168.13.1371.
- Rothman, K. J., S. Greenland, and T. L. Lash. 2008. *Modern Epidemiology* 3rd Ed., Philadelphia: Lippincott Williams & Wilkins.

- Rothwell, P. M. 2005. "External Validity of Randomised Controlled Trials: 'To Whom Do the Results of This Trial Apply?'" *Lancet*, 365(9453): 82–93, doi:10.1016/S0140-6736(04)17670-8.
- Rubin, D. B. 1978. "Bayesian Inference for Causal Effects: The Role of Randomization." *The Annals of Statistics*, 6(1): 34–58.
- Rubin, D. B. 2008. "For Objective Causal Inference, Design Trumps Analysis." *The Annals of Applied Statistics*, 2(3): 808–840, doi:10.1214/08-AOAS187.
- Rubin, D. B. 2010. "On the Limitations of Comparative Effectiveness Research." *Statistics in Medicine*, 29(19): 1991–1995, doi:10.1002/sim.3960.
- Rubin, D. B. 2007. "The Design versus the Analysis of Observational Studies for Causal Effects: Parallels with the Design of Randomized Trials." *Statistics in Medicine*, 26(1): 20–36, doi:10.1002/sim.2739.
- Sanders, C., A. Rogers, R. Bowen, P. Bower, S. Hirani, M. Cartwright, R. Fitzpatrick, M. Knapp, J. Barlow, J. Hendy, T. Chrysanthaki, M. Bardsley, and S. P. Newman. 2012. "Exploring Barriers to Participation and Adoption of Telehealth and Telecare within the Whole System Demonstrator Trial: A Qualitative Study." *BMC Health Services Research*, 12: 220, doi:10.1186/1472-6963-12-220.
- Schoenfeld, D. 1982. "Partial Residuals for the Proportional Hazards Regression Model." *Biometrika*, 69(1): 239–241.
- Schulz, K. F., D. G. Altman, and D. Moher. 2010. "CONSORT 2010 Statement: Updated Guidelines for Reporting Parallel Group Randomised Trials." *BMJ*, 340: 698–702, doi:10.1136/bmj.c332.
- Schwartz, D. and J. Lellouch. 1967. "Explanatory and Pragmatic Attitudes in Therapeutical Trials." *Journal of Chronic Diseases*, 20(8): 637–648.
- Scriven, M. 1966. *The Methodology of Evaluation*, Lafayette: Social Science Education Consortium, Purdue University.
- Sculpher, M. J., F. S. Pang, A. Manca, M. F. Drummond, S. Golder, H. Urdahl, L. M. Davies, and A. Eastwood. 2004. "Generalisability in Economic Evaluation Studies in Healthcare: A Review and Case Studies." *Health Technology Assessment*, 8(49): iii–iv, 1–192.
- Sekhon, J. S. and R. D. Grieve. 2012. "A Matching Method for Improving Covariate Balance in Cost-Effectiveness Analyses." *Health Economics*, 21(6): 695–714, doi:10.1002/hec.1748.

- Shadish, W. R., M. H. Clark, and P. M. Steiner. 2008. "Can Nonrandomized Experiments Yield Accurate Answers? A Randomized Experiment Comparing Random and Nonrandom Assignments." *Journal of the American Statistical Association*, 103(484): 1334–1343, doi:10.1198/016214508000000733.
- Shea, S., R. S. Weinstock, J. A. Teresi, W. Palmas, J. Starren, J. J. Cimino, A. M. Lai, L. Field, P. C. Morin, R. Goland, R. E. Izquierdo, S. Ebner, S. Silver, E. Petkova, J. Kong, and J. P. Eimicke. 2009. "A Randomized Trial Comparing Telemedicine Case Management with Usual Care in Older, Ethnically Diverse, Medically Underserved Patients with Diabetes Mellitus: 5 Year Results of the IDEATel Study." *Journal of the American Medical Informatics Association*, 16(4): 446–456, doi:10.1197/jamia.M3157.
- Sheikh, A., T. Cornford, N. Barber, A. Avery, A. Takian, V. Lichtner, D. Petrakaki, S. Crowe, K. Marsden, A. Robertson, Z. Morrison, E. Klecun, R. Prescott, C. Quinn, Y. Jani, M. Ficociello, K. Voutsina, J. Paton, B. Fernando, A. Jacklin, and K. Cresswell. 2011. "Implementation and Adoption of Nationwide Electronic Health Records in Secondary Care in England: Final Qualitative Results from Prospective National Evaluation in 'Early Adopter' Hospitals." *BMJ*, 343: d6054, doi:10.1136/bmj.d6054.
- Shrank, W. 2013. "The Center For Medicare And Medicaid Innovation's Blueprint for Rapid-Cycle Evaluation of New Care and Payment Models." *Health Affairs*, 32(4): 807–812, doi:10.1377/hlthaff.2013.0216.
- Sicotte, C., G. Paré, S. Morin, J. Potvin, and M.-P. Moreault. 2011. "Effects of Home Telemonitoring to Support Improved Care for Chronic Obstructive Pulmonary Diseases." *Telemedicine and e-Health*, 17(2): 95–103, doi:10.1089/tmj.2010.0142.
- Snowden, D. J. and M. E. Boone. 2007. "A Leader's Framework for Decision Making. A Leader's Framework for Decision Making." *Harvard Business Review*, 85(11): 68–76, 149.
- Sohn, S., T. M. Helms, J. T. Pelleter, A. Müller, A. I. Kröttinger, and O. Schöffski. 2012. "Costs and Benefits of Personalized Healthcare for Patients with Chronic Heart Failure in the Care and Education Program 'Telemedicine for the Heart?'" *Telemedicine and e-Health*, 18(3): 198–204, doi:10.1089/tmj.2011.0134.
- Sood, S., V. Mbarika, S. Jugoo, R. Dookhy, C. R. Doarn, N. Prakash, and R. C. Merrell. 2007. "What Is Telemedicine? A Collection of 104 Peer-Reviewed Perspectives and Theoretical Underpinnings." *Telemedicine and e-Health*, 13(5): 573–590, doi:10.1089/tmj.2006.0073.

- Spencer, S. A. and M. P. Davies. 2012. "Hospital Episode Statistics: Improving the Quality and Value of Hospital Data: A National Internet E-Survey of Hospital Consultants." *BMJ Open*, 2: e001651, doi:10.1136/bmjopen-2012-001651.
- Spiegelhalter, D. J., J. P. Myles, D. R. Jones, and K. R. Abrams. 1999. "An Introduction to Bayesian Methods in Health Technology Assessment." *BMJ*, 319(7208): 508–512.
- Splawa-Neyman, J., D. M. Dabrowska, and T. P. Speed. 1990. "On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9." *Statistical Science*, 5(4): 465–472.
- Van Staa, T.-P., L. Dyson, G. McCann, S. Padmanabhan, R. Belatri, B. Goldacre, J. Cassell, M. Pirmohamed, D. Torgerson, S. Ronaldson, J. Adamson, A. Taweel, B. Delaney, S. Mahmood, S. Baracaia, T. Round, R. Fox, T. Hunter, M. Gulliford, and L. Smeeth. 2014. "The Opportunities and Challenges of Pragmatic Point-of-Care Randomised Trials Using Routinely Collected Electronic Records: Evaluations of Two Exemplar Trials." *Health Technology Assessment*, 18(43): 1–146, doi:10.3310/hta18430.
- Van Staa, T.-P., B. Goldacre, M. Gulliford, J. Cassell, M. Pirmohamed, A. Taweel, B. Delaney, and L. Smeeth. 2012. "Pragmatic Randomised Trials Using Routine Electronic Health Records: Putting Them to the Test." *BMJ*, 344: e55, doi:10.1136/bmj.e55.
- StataCorp. 2009. Stata Statistical Software: Release 11, College Station, Texas: StataCorp LP.
- De Stavola, B. L., R. M. Daniel, G. B. Ploubidis, and N. Micali. 2015. "Mediation Analysis with Intermediate Confounding: Structural Equation Modeling Viewed through the Causal Inference Lens." *American Journal of Epidemiology*, 181(1): 64–80, doi:10.1093/aje/kwu239.
- Steventon, A. 2013. "Making the Best Use of Administrative Data." *BMJ*, 346: f1284, doi:10.1136/bmj.f1284.
- Steventon, A., M. Bardsley, J. Billings, J. Dixon, H. Doll, M. Beynon, S. Hirani, M. Cartwright, L. Rixon, M. Knapp, C. Henderson, A. Rogers, J. Hendy, R. Fitzpatrick, and S. Newman. 2013a. "Effect of Telecare on Use of Health and Social Care Services: Findings from the Whole Systems Demonstrator Cluster Randomised Trial." *Age and Ageing*, 42(4): 501–508, doi:10.1093/ageing/aft008.
- Steventon, A., M. Bardsley, J. Billings, J. Dixon, H. Doll, S. Hirani, M. Cartwright, L. Rixon, M. Knapp, C. Henderson, A. Rogers, R. Fitzpatrick, J. Hendy, and S. Newman. 2012a. "Effect of Telehealth on Use of Secondary Care and Mortality: Findings from the Whole System Demonstrator Cluster Randomised Trial." *BMJ*, 344: e3874, doi:10.1136/bmj.e3874.

- Steventon, A., M. Bardsley, J. Billings, T. Georghiou, and G. Lewis. 2011. An Evaluation of the Impact of Community-Based Interventions on Hospital Use: A Case Study of Eight Partnership for Older People Projects, London: Nuffield Trust.
- Steventon, A., M. Bardsley, J. Billings, T. Georghiou, and G. H. Lewis. 2012b. “The Role of Matched Controls in Building an Evidence Base for Hospital-Avoidance Schemes: A Retrospective Evaluation.” *Health Services Research*, 47(4): 1679–1698, doi:10.1111/j.1475-6773.2011.01367.x.
- Steventon, A., M. Bardsley, H. Doll, E. Tuckey, and S. P. Newman. 2014. “Effect of Telehealth on Glycaemic Control: Analysis of Patients with Type 2 Diabetes in the Whole Systems Demonstrator Cluster Randomised Trial.” *BMC Health Services Research*, 14: 334, doi:10.1186/1472-6963-14-334.
- Steventon, A., M. Bardsley, and N. Mays. 2014. “Effect of a Telephonic Alert System (Healthy Outlook) for Patients with Chronic Obstructive Pulmonary Disease: Cohort Study with Matched Controls.” *Journal of Public Health*. Epub Ahead, doi:10.1093/pubmed/fdu042.
- Steventon, A., R. Grieve, and M. Bardsley. “An Approach to Assess Generalizability in Comparative Effectiveness Research: A Case Study of the Whole Systems Demonstrator Cluster Randomized Trial Comparing Telehealth with Usual Care for Patients with Chronic Health Conditions.” *Medical Decision Making*.
- Steventon, A., R. Grieve, and J. S. Sekhon. 2015. “A Comparison of Alternative Strategies for Choosing Control Populations in Observational Studies.” *Health Services and Outcomes Research Methodology*, Epub Ahead, doi:10.1007/s10742-014-0135-8.
- Steventon, A., S. Tunkel, I. Blunt, and M. Bardsley. 2013b. “Effect of Telephone Health Coaching (Birmingham OwnHealth) on Hospital Use and Associated Costs: Cohort Study with Matched Controls.” *BMJ*, 347: f4585, doi:10.1136/bmj.f4585.
- Stoddart, A., J. Hanley, S. Wild, C. Pagliari, M. Paterson, S. Lewis, A. Sheikh, A. Krishan, P. Padfield, and B. McKinstry. 2013. “Telemonitoring-Based Service Redesign for the Management of Uncontrolled Hypertension (HITS): Cost and Cost-Effectiveness Analysis of a Randomised Controlled Trial.” *BMJ Open*, 3(5): e002681, doi:10.1136/bmjopen-2013-002681.
- Stuart, E. A. 2010. “Matching Methods for Causal Inference: A Review and a Look Forward.” *Statistical Science*, 25(1): 1–21, doi:10.1214/09-STS313.
- Stuart, E. A., S. R. Cole, C. P. Bradshaw, and P. J. Leaf. 2011. “The Use of Propensity Scores to Assess the Generalizability of Results from Randomized Trials.” *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 174(2): 369–386, doi:10.1111/j.1467-985X.2010.00673.x.

- Stuart, E. A., E. DuGoff, M. Abrams, D. Salkever, and D. Steinwachs. 2013. “Estimating Causal Effects in Observational Studies Using Electronic Health Data: Challenges and (some) Solutions.” *eGEMs (Generating Evidence & Methods to improve patient outcomes)*, 1(3): 4, doi:10.13063/2327-9214.1038.
- Stuart, E. A., H. A. Huskamp, K. Duckworth, J. Simmons, Z. Song, M. E. Chernew, and C. L. Barry. 2014. “Using Propensity Scores in Difference-in-Differences Models to Estimate the Effects of a Policy Change.” *Health Services and Outcomes Research Methodology*, 14(4): 166–182, doi:10.1007/s10742-014-0123-z.
- Stuart, E. A. and D. B. Rubin. 2008. “Matching with Multiple Control Groups with Adjustment for Group Differences.” *Journal of Educational and Behavioral Statistics*, 33(3): 279–306, doi:10.3102/1076998607306078.
- Stukel, T. A., E. S. Fisher, D. E. Wennberg, D. A. Alter, D. J. Gottlieb, and M. J. Vermeulen. 2007. “Analysis of Observational Studies in the Presence of Treatment Selection Bias: Effects of Invasive Cardiac Management on AMI Survival Using Propensity Score and Instrumental Variable Methods.” *JAMA*, 297(3): 278–285.
- Stürmer, T., M. Joshi, R. J. Glynn, J. Avorn, K. J. Rothman, and S. Schneeweiss. 2006. “A Review of the Application of Propensity Score Methods Yielded Increasing Use, Advantages in Specific Settings, but Not Substantially Different Estimates Compared with Conventional Multivariable Methods.” *Journal of Clinical Epidemiology*, 59(5): 437–447, doi:10.1016/j.jclinepi.2005.07.004.
- Subramanian, U., F. Hopp, J. Lowery, P. Woodbridge, and D. Smith. 2004. “Research in Home-Care Telemedicine: Challenges in Patient Recruitment.” *Telemedicine Journal and e-Health*, 10(2): 155–161, doi:10.1089/tmj.2004.10.155.
- Takahashi, P. Y., G. J. Hanson, J. L. Pecina, R. J. Stroebel, R. Chaudhry, N. D. Shah, and J. M. Naessens. 2010. “A Randomized Controlled Trial of Telemonitoring in Older Adults with Multiple Chronic Conditions: The Tele-ERA Study.” *BMC Health Services Research*, 10: 255, doi:10.1186/1472-6963-10-255.
- Takahashi, P. Y., J. L. Pecina, B. Upatising, R. Chaudhry, N. D. Shah, H. Van Houten, S. Cha, I. Croghan, J. M. Naessens, and G. J. Hanson. 2012. “A Randomized Controlled Trial of Telemonitoring in Older Adults With Multiple Health Issues to Prevent Hospitalizations and Emergency Department Visits.” *JAMA Internal Medicine*, 172(10): 773–779, doi:10.1001/archinternmed.2012.256.

- Tan, S. Y., L. L. Low, Y. Yang, and K. H. Lee. 2013. "Applicability of a Previously Validated Readmission Predictive Index in Medical Patients in Singapore: A Retrospective Study." *BMC Health Services Research*, 13: 366, doi:10.1186/1472-6963-13-366.
- The Health & Social Care Information Centre. 2013. "Hospital Episode Statistics," Available at: <http://www.hscic.gov.uk/hes>.
- The NHS Information Centre for health and social care. 2012. Quality and Outcomes Framework, Achievement, Prevalence and Exceptions Data, 2011/12, Leeds: Health and Social Care Information Centre.
- Thompson, S. G. and J. A. Barber. 2000. "How Should Cost Data in Pragmatic Randomised Trials Be Analysed?" *BMJ*, 320(7243): 1197–1200.
- Trief, P. M., J. A. Teresi, J. P. Eimicke, S. Shea, and R. S. Weinstock. 2009. "Improvement in Diabetes Self-Efficacy and Glycaemic Control Using Telemedicine in a Sample of Older, Ethnically Diverse Individuals Who Have Diabetes: The IDEATel Project." *Age and Ageing*, 38(2): 219–225, doi:10.1093/ageing/afn299.
- Tsai, T. C., K. E. Joynt, E. J. Orav, A. A. Gawande, and A. K. Jha. 2013. "Variation in Surgical-Readmission Rates and Quality of Hospital Care." *New England Journal of Medicine*, 369(12): 1134–1142, doi:10.1056/NEJMsa1303118.
- Vuorinen, A.-L., J. Leppänen, H. Kaijanranta, M. Kulju, T. Heliö, M. van Gils, and J. Lähteenmäki. 2014. "Use of Home Telemonitoring to Support Multidisciplinary Care of Heart Failure Patients in Finland: Randomized Controlled Trial." *Journal of Medical Internet Research*, 16(12): e282, doi:10.2196/jmir.3651.
- Wagner, E. H. 1998. "Chronic Disease Management: What Will It Take to Improve Care for Chronic Illness?" *Effective Clinical Practice*, 1(1): 2–4.
- Wallihan, D. B., T. E. Stump, and C. M. Callahan. 1999. "Accuracy of Self-Reported Health Services Use and Patterns of Care among Urban Older Adults." *Medical Care*, 37(7): 662–670.
- Van Walraven, C. and P. Austin. 2012. "Administrative Database Research Has Unique Characteristics That Can Risk Biased Results." *Journal of Clinical Epidemiology*, 65(2): 126–131, doi:10.1016/j.jclinepi.2011.08.002.
- Wanless, D., J. Forder, J.-L. Fernandez, L. Beesley, M. Henwood, and F. Moscone. 2006. Securing Good Care for Older People: Taking a Long-Term View, London: King's Fund.

- Ward, E., M. King, M. Lloyd, P. Bower, and K. Friedl. 1999. "Conducting Randomized Trials in General Practice: Methodological and Practical Issues." *British Journal of General Practice*, 49: 919–922.
- Welschen, L. M. C., E. Bloemendal, G. Nijpels, J. M. Dekker, R. J. Heine, W. A. B. Stalman, and L. M. Bouter. 2005. "Self-Monitoring of Blood Glucose in Patients with Type 2 Diabetes Mellitus Who Are Not Using Insulin (Review)." *Cochrane Database of Systematic Reviews*, 2: CD005060.
- Wennberg, D. E., A. Marr, L. Lang, S. O'Malley, and G. Bennett. 2010. "A Randomized Trial of a Telephone Care-Management Strategy." *The New England Journal of Medicine*, 363(13): 1245–1255, doi:10.1056/NEJMsa0902321.
- Wennberg, D. E., S. M. Sharp, G. Bevan, J. S. Skinner, D. J. Gottlieb, and J. E. Wennberg. 2014. "A Population Health Approach to Reducing Observational Intensity Bias in Health Risk Adjustment: Cross Sectional Analysis of Insurance Claims." *BMJ*, 348: g2392, doi:10.1136/bmj.g2392.
- Wennberg, D., M. Siegel, B. Darin, N. Filipova, R. Russell, L. Kenney, K. Steinort, T.-R. Park, G. Cakmakci, J. Dixon, N. Curry, and J. Billings. 2006. Combined Predictive Model: Final Report and Technical Documentation, London: King's Fund.
- Wennberg, J. E., E. S. Fisher, and J. S. Skinner. 2002. "Geography And The Debate Over Medicare Reform." *Health Affairs*, 23(Suppl): W96–W114, doi:10.1377/hlthaff.w2.96.
- Wennberg, J. E., D. O. Staiger, S. M. Sharp, D. J. Gottlieb, G. Bevan, K. McPherson, and H. G. Welch. 2013. "Observational Intensity Bias Associated with Illness Adjustment: Cross Sectional Analysis of Insurance Claims." *BMJ*, 346: f549, doi:10.1136/bmj.f549.
- West, S. G., N. Duan, W. Pequegnat, P. Gaist, D. C. Des Jarlais, D. Holtgrave, J. Szapocznik, M. Fishbein, B. Rapkin, M. Clatts, and P. D. Mullen. 2008. "Alternatives to the Randomized Controlled Trial." *American Journal of Public Health*, 98(8): 1359–1366, doi:10.2105/AJPH.2007.124446.
- Wherton, J., P. Sugarhood, R. Procter, M. Rouncefield, G. Dewsbury, S. Hinder, and T. Greenhalgh. 2012. "Designing Assisted Living Technologies 'in the Wild': Preliminary Experiences with Cultural Probe Methodology." *BMC Medical Research Methodology*, 12: 188, doi:10.1186/1471-2288-12-188.
- Windle, K., R. Wagland, J. Forder, F. D'Amico, D. Janssen, and G. Wistow. 2009. The National Evaluation of Partnerships for Older People Projects: Executive Summary, Canterbury, UK: Personal Social Services Research Unit, University of Kent.

- Woollard, M. 2014. "Administrative Data: Problems and Benefits: A Perspective from the United Kingdom." In *Facing the Future: European Research Infrastructures for the Humanities and Social Sciences*, edited by A. Duşa, D. Nelle, G. Stock, and G. G. Wagner, pp. 49–60. Berlin: SCIVERO Verlag.
- Xu, W., J.-P. Collet, S. Shapiro, Y. Lin, T. Yang, R. W. Platt, C. Wang, and J. Bourbeau. 2008. "Independent Effect of Depression and Anxiety on Chronic Obstructive Pulmonary Disease Exacerbations and Hospitalizations." *American Journal of Respiratory and Critical Care Medicine*, 178(9): 913–20, doi:10.1164/rccm.200804-619OC.
- Yudkin, P. L. and C. W. G. Redman. 1990. "Obstetric Audit Using Routinely Collected Computerised Data." *BMJ*, 301(6765): 1371–1373.
- Zelen, M. 1990. "Randomized Consent Designs for Clinical Trials: An Update." *Statistics in Medicine*, 9(6): 645–656, doi:10.1002/sim.4780090611.

Appendix: R code for simulations

```
library('MASS')
```

```
library('Matching')
```

```
library('lme4')
```

```
#####
```

```
##### constants
```

```
#####
```

```
# number of simulations
```

```
NSIM = 20000
```

```
# number of control areas (i.e., excluding intervention area)
```

```
NAREA = 49
```

```
# number of people in each area
```

```
NPOP = 1000
```

```
# number of iterations to perform to produce target saturation (set to a high number)
```

```
NSAT = 1000000
```

```
#####  
##### run code for the various sensitivity analyses considered in the paper  
#####
```

```
# compile all the functions below before running these!
```

```
main(N_run = 0.1, a2_run = 0.3)
```

```
main(N_run = 0.3, a2_run = 0.3)
```

```
main(N_run = 0.5, a2_run = 0.3)
```

```
main(N_run = 0.1, a2_run = 0.1)
```

```
main(N_run = 0.3, a2_run = 0.1)
```

```
main(N_run = 0.5, a2_run = 0.1)
```

```
main(N_run = 0.1, a2_run = 0.2)
```

```
main(N_run = 0.3, a2_run = 0.2)
```

```
main(N_run = 0.5, a2_run = 0.2)
```

```
#####  
##### functions follow below  
#####
```

```
# this is the main function that outputs arrays of balance statistics and estimated treatment effects
```

```

main <- function(N_run, a2_run) {

#####

#### set assumptions

#####

# set mean of area-level variables

amean <- c(0,0,0)

# set covariance matrix of x1,1 and x2,1

icov <- diag(2)
icov[1,2] <- 0.2
icov[2,1] <- 0.2

# set propensity score coefficients

a <- c(-1.7, 0.5, a2_run)

# set target saturation

N <- N_run

```

```

#####

##### calibrate propensity score

#####

# function returns saturation for given covariate

find_sat <- function(x) {

  avar_m1 <- c(1,1,1)

  ivar_m1 <- mvrnorm(n=NSAT, mu=c(1, avar_m1[2]), Sigma=icov)

  p_m1 <- x + a[2]*ivar_m1[,1] + a[3]*ivar_m1[,2]

  p_m1 <- exp(p_m1) / (1+exp(p_m1))

  tr_m1 <- rbinom(NSAT, 1, p_m1)

  mean(tr_m1)-N

}

# use R optimisation routine to find the covariate value that produces target saturation in
expectation

```

```
opt <- uniroot(find_sat, c(-10,10))
```

```
a[1] <- opt$root
```

```
#####
```

```
##### conduct simulations
```

```
#####
```

```
# this will contain estimated treatment effects
```

```
r_c <- NULL
```

```
# this will contain balance statistics (standardised differences)
```

```
b_c <- NULL
```

```
# loop through iterations of the data generating process
```

```
for(i in 1:NSIM) {
```

```
#####
```

```

##### obtain matched data from all strategies

#####

m <- getpairs(i, a, amean, icov)

# get number of matched intervention patients

n_matched <- sum(1*(m$tr==1)*(m$meth=="Method 1"))

# get balance statistics (standardised differences) under each strategy

b1 <- MatchBalance(tr~x1+x2, data = subset(m,m$meth=="Method 1"), ks =
FALSE, nboots = 0, print.level = 0)

b2 <- MatchBalance(tr~x1+x2, data = subset(m,m$meth=="Method 2"), ks =
FALSE, nboots = 0, print.level = 0)

b3 <- MatchBalance(tr~x1+x2, data = subset(m,m$meth=="Method 3"), ks =
FALSE, nboots = 0, print.level = 0)

b4 <- MatchBalance(tr~x1+x2, data = subset(m,m$meth=="Method 4"), ks =
FALSE, nboots = 0, print.level = 0)

# sort these out into a tidy array and add to b_c

compile <- function(df,meth) {

  var <- c("x1", "x2")

  sdiff <- c(df$BeforeMatching[[1]]$sdiff, df$BeforeMatching[[2]]$sdiff)

```

```

        meanTr <-
c(df$BeforeMatching[[1]]$mean.Tr,df$BeforeMatching[[2]]$mean.Tr)

        meanCo <- c(df$BeforeMatching[[1]]$mean.Co,
df$BeforeMatching[[2]]$mean.Co)

        varTr <- c(df$BeforeMatching[[1]]$var.Tr, df$BeforeMatching[[2]]$var.Tr)
        varCo <- c(df$BeforeMatching[[1]]$var.Co, df$BeforeMatching[[2]]$var.Co)

        data.frame(var, sdiff, meanTr, meanCo, varTr, varCo, factor(meth),
n_matched)

    }

```

```

    b <- rbind(compile(b1, "Method 1"), compile(b2, "Method 2"), compile(b3,
"Method 3"), compile(b4, "Method 4"))

```

```

    b$sim <- i

```

```

    b_c <- rbind(b_c,b)

```

```

#####

```

```

##### get estimated treatment effects under each scenario (for both binary and
normally-distributed outcome)

```

```

#####

```

```

r1 <- getresponse(

```

```

    m=m,

```

```

    b0 = 1.7,

```

```

    b1 = 0.3,

```

```
b2 = 0.15,  
b3 = 0.01,  
b4 = 0.05,  
b5 = 0.06,  
delta = 0,  
scen = "Base case",  
i=i)
```

```
r2 <- getresponse(  
  m=m,  
  b0 = 1.7,  
  b1 = 0.3,  
  b2 = 0,  
  b3 = 0,  
  b4 = 0,  
  b5 = 0,  
  delta = 0,  
  scen = "Simple confounding",  
  i=i)
```

```
r3 <- getresponse(  
  m=m,  
  b0 = 1.7,  
  b1 = 0.3,
```

```
b2 = 0.15,  
b3 = 0,  
b4 = 0,  
b5 = 0,  
delta = 0,  
scen = "No area-level variation",  
i=i)
```

```
r4 <- getresponse(  
  m=m,  
  b0 = 1.7,  
  b1 = 0.3,  
  b2 = 0.15,  
  b3 = 0.01,  
  b4 = 0.05,  
  b5 = 0,  
  delta = 0,  
  scen = "No unexplained area-level variation",  
  i=i)
```

```
r5 <- getresponse(  
  m=m,  
  b0 = 1.7,  
  b1 = 0.3,
```

```
b2 = 0.15,  
b3 = 0.01,  
b4 = 0.05,  
b5 = 0.3,  
delta = 0,  
scen = "High unexplained area-level variation",  
i=i)
```

```
# combine estimated treatment effects from all scenarios
```

```
r <- rbind(r1, r2, r3, r4, r5)
```

```
r_c <- rbind(r_c, r)
```

```
}
```

```
#####
```

```
# save output files
```

```
#####
```

```
desc <- paste("N(",N,") ", "a2(", a[3], ")",
```

```
sep = "")
```

```

path <- "C:/" # update path for output files as appropriate

filename_r <- paste(path,"Response ", desc, ".csv", sep="")
filename_b <- paste(path,"Balance ", desc, ".csv", sep="")
filename_mor <- paste(path,"MOR ", desc, ".csv", sep="")

write.csv(r_c, filename_r)
write.csv(b_c, filename_b)
write.csv(estmor_c, filename_mor)

}

#####

##### this function simulates outcome data and returns estimated treatment effects

#####

getresponse <- function(m, b0, b1, b2, b3, b4, b5, delta, scen, i) {

```

```
#####

#### first assume binary outcome

#####

# simulate outcomes with actual treatment assignment

# note that, in the notation of the paper, b1-b5 are b1,1, b2,1, b1,2, b2,2, and b3,2,
respectively

logit <- b0 + b1*m$x1 + b2*m$x2 + b3*m$x3 + b4*m$x4 + b5*m$x5 + delta*m$tr

logit <- exp(logit) / (1+exp(logit))

n <- nrow(m)

y <- rbinom(n, 1, logit)

# estimate treatment effects (using difference-in-means estimator)

eff_m1 <- mean(y[m$meth=="Method 1" & m$tr ==1]) - mean(y[m$meth=="Method 1"
& m$tr ==0])

eff_m2 <- mean(y[m$meth=="Method 2" & m$tr ==1]) - mean(y[m$meth=="Method 2"
& m$tr ==0])

eff_m3 <- mean(y[m$meth=="Method 3" & m$tr ==1]) - mean(y[m$meth=="Method 3"
& m$tr ==0])

eff_m4 <- mean(y[m$meth=="Method 4" & m$tr ==1]) - mean(y[m$meth=="Method 4"
& m$tr ==0])

meth <- c("Method 1", "Method 2", "Method 3", "Method 4")
```

```

eff <- c(eff_m1, eff_m2, eff_m3, eff_m4)

#####

##### then with normally-distributed outcome

#####

# simulate outcomes with actual treatment assignment

linear <- b0 + b1*m$x1 + b2*m$x2 + b3*m$x3 + b4*m$x4 + b5*m$x5 + delta*m$tr

n <- nrow(m)

y <- rnorm(n, linear, sd = 0.5)

# estimate treatment effects (using difference-in-means estimator)

eff_m1 <- mean(y[m$meth=="Method 1" & m$tr ==1]) - mean(y[m$meth=="Method 1"
& m$tr ==0])

eff_m2 <- mean(y[m$meth=="Method 2" & m$tr ==1]) - mean(y[m$meth=="Method 2"
& m$tr ==0])

eff_m3 <- mean(y[m$meth=="Method 3" & m$tr ==1]) - mean(y[m$meth=="Method 3"
& m$tr ==0])

eff_m4 <- mean(y[m$meth=="Method 4" & m$tr ==1]) - mean(y[m$meth=="Method 4"
& m$tr ==0])

```

```
meth <- c("Method 1", "Method 2", "Method 3", "Method 4")
```

```
eff_norm <- c(eff_m1, eff_m2, eff_m3, eff_m4)
```

```
#####
```

```
##### finally return estimated treatment effects to main loop
```

```
#####
```

```
data.frame(
```

```
  meth = c(meth, meth),
```

```
  type = c(rep("Binary",4), rep("Normal",4)),
```

```
  eff = c(eff, eff_norm),
```

```
  scen = scen,
```

```
  sim = i)
```

```
}
```

```
#####
```

```
##### this function produces matched pairs from all strategies
```

```
#####
```

```
getpairs <- function(i, a, amean, icov) {
```

```

# generate area-level variables (x1,2-x3,2)

# (assumed independent)

avar <- mvrnorm(n=NAREA, mu=amean, Sigma=diag(3))

# select matched control area by minimizing Euclidean distance from the intervention area

# note values for intervention area are hardcoded as 1

t <- (avar-1)*(avar-1)

euclid <- t[,1] + t[,2]

match <- which(euclid == min(euclid))

# select random control area

area.random <- ceiling(runif(1, 0, NAREA-0.00001))

# compile area-level covariates for each strategy

# strategy 4 (national controls) is dealt with below

avar_m1 <- c(1,1,1) # strategy 1 (local controls)

avar_m2 <- avar[area.random,] # strategy 2 (random area)

avar_m3 <- avar[match,] # strategy 3 (matched area)

```

```

# generate individual-level covariates for all individuals in intervention area

ivar_m1 <- mvrnorm(n=NPOP, mu=c(1, avar_m1[2]), Sigma=icov)

# slightly more difficult for other three methods as we need to do all areas

ivar <- array(0, dim=c(NAREA,NPOP,2))

for(j in 1:NAREA) {
  ivar[j,,] <- mvrnorm(n=NPOP, mu=c(1, avar[j,2]), Sigma=icov)
}

ivar_m2 <- ivar[area.random,,] # strategy 2
(random area)

ivar_m3 <- ivar[match,,] # strategy 3
(matched area)

ivar_m4 <- cbind(c(ivar[,1]), c(ivar[,2])) # strategy 4 (national)

# generate true propensity score

p_m1 <- a[1] + a[2]*ivar_m1[,1] + a[3]*ivar_m1[,2]

p_m1 <- exp(p_m1) / (1+exp(p_m1))

# these arrays will hold treatment assignments

```

```

ab <- NPOP*NAREA

tr_m1 <- rbinom(NPOP, 1, p_m1)

tr_m2 <- rep(0,NPOP)

tr_m3 <- rep(0,NPOP)

tr_m4 <- rep(0,ab)

# tidy up data by putting it into data frames (not strictly necessary)

# this is slightly more difficult for method 4 (national), as area-level covariates are not
constant

data_m1 <- data.frame(tr = tr_m1, x1 = ivar_m1[,1], x2 = ivar_m1[,2], x3 = avar_m1[1], x4
= avar_m1[2], x5 = avar_m1[3])

data_m2 <- data.frame(tr = tr_m2, x1 = ivar_m2[,1], x2 = ivar_m2[,2], x3 = avar_m2[1], x4
= avar_m2[2], x5 = avar_m2[3])

data_m3 <- data.frame(tr = tr_m3, x1 = ivar_m3[,1], x2 = ivar_m3[,2], x3 = avar_m3[1], x4
= avar_m3[2], x5 = avar_m3[3])

data_m4 <- data.frame(

  tr = tr_m4,

  x1 = ivar_m4[,1],

  x2 = ivar_m4[,2],

  x3 = avar[ceiling((1:(NPOP*NAREA))/NPOP),1],

  x4 = avar[ceiling((1:(NPOP*NAREA))/NPOP),2],

  x5 = avar[ceiling((1:(NPOP*NAREA))/NPOP),3]

)

```

```
# fit empirical propensity score to data from intervention area, then apply coefficients to
other areas
```

```
fm <- glm(tr~x1, data = data_m1, family=binomial(link="logit"))
```

```
data_m1$p_est <- fm$coeff[1] + fm$coeff[2] * data_m1$x1
```

```
data_m2$p_est <- fm$coeff[1] + fm$coeff[2] * data_m2$x1
```

```
data_m3$p_est <- fm$coeff[1] + fm$coeff[2] * data_m3$x1
```

```
data_m4$p_est <- fm$coeff[1] + fm$coeff[2] * data_m4$x1
```

```
data_m1$p_est <- exp(data_m1$p_est) / (1+exp(data_m1$p_est))
```

```
data_m2$p_est <- exp(data_m2$p_est) / (1+exp(data_m2$p_est))
```

```
data_m3$p_est <- exp(data_m3$p_est) / (1+exp(data_m3$p_est))
```

```
data_m4$p_est <- exp(data_m4$p_est) / (1+exp(data_m4$p_est))
```

```
#####
```

```
# now we will form matched pairs
```

```
#####
```

```
# select cases from intervention area
```

```
cases <- subset(data_m1,data_m1$tr==1)
```

```

# append cases to data frames for the three external strategies

data_m2_app <- rbind(data_m2,cases)

data_m3_app <- rbind(data_m3,cases)

data_m4_app <- rbind(data_m4,cases)

# do matching on 1-1 basis with replacement

m1 <- Match(Tr=data_m1$tr, X=data_m1$p_est, M=1, estimand = "ATT", replace =
TRUE, ties = FALSE)

m2 <- Match(Tr=data_m2_app$tr, X=data_m2_app$p_est, M=1, estimand = "ATT",
replace = TRUE, ties = FALSE)

m3 <- Match(Tr=data_m3_app$tr, X=data_m3_app$p_est, M=1, estimand = "ATT",
replace = TRUE, ties = FALSE)

m4 <- Match(Tr=data_m4_app$tr, X=data_m4_app$p_est, M=1, estimand = "ATT",
replace = TRUE, ties = FALSE)

# compile matched data sets under each strategy

match_m1 <- data_m1[c(m1$index.treated, m1$index.control),]

match_m2 <- data_m2_app[c(m2$index.treated, m2$index.control),]

match_m3 <- data_m3_app[c(m3$index.treated, m3$index.control),]

match_m4 <- data_m4_app[c(m4$index.treated, m4$index.control),]

# add pair ID

```

```

match_m1$pair <- rep(1:length(m1$index.treated),2)

match_m2$pair <- rep(1:length(m2$index.treated),2)

match_m3$pair <- rep(1:length(m3$index.treated),2)

match_m4$pair <- rep(1:length(m4$index.treated),2)

# produce combined data set of matched data from all strategies

match_m1$meth <- factor("Method 1")

match_m2$meth <- factor("Method 2")

match_m3$meth <- factor("Method 3")

match_m4$meth <- factor("Method 4")

match_all <- rbind(match_m1, match_m2, match_m3, match_m4)

match_all$sim <- i

# return combined data set to main loop

match_all
}

```