ECONOMIC EVALUATION OF CHANGES TO THE ORGANISATION AND DELIVERY OF HEALTH SERVICES: METHODS AND APPLICATIONS

A thesis submitted to the University of Manchester for the degree of Doctor of Philosophy (PhD) in the Faculty of Biology, Medicine and Health

2017

Rachel L Meacock

School of Health Sciences

Contents

Word count: 49,236

List of tables

List of figures

List of abbreviations

A&E	Accident and Emergency
AFT	Accelerated failure time
AIC	Akaike Information Criterion
AMI	Acute myocardial infarction
AQ	Advancing Quality
CABG	Coronary artery bypass grafting
CQUIN	Commissioning for Quality and Innovation
CUA	Cost-utility analysis
DANQALE	Discounted and quality-adjusted life expectancy
DiD	Difference-in-differences
DRG	Diagnostic-related group
GP	General practitioner
HQID	Hospital Quality Incentive Demonstration
HRG	Healthcare resource group
HSCIC	Health and Social Care Information Centre
HTA	Health technology assessment
ICD	International classification of disease
ICER	Incremental cost-effectiveness ratio
LOS	Length of stay
LSOA	Lower-layer super output area
MRC	Medical Research Council
NHS	National Health Service
NICE	National Institute for Health and Care Excellence
OLS	Ordinary least squares
ONS	Office for National Statistics
OR	Odds ratio
P4P	Pay-for-performance
PROMs	Patient-reported outcome measures
PSS	Personal social services
QALY	Quality-adjusted life year
QOF	Quality and Outcomes Framework
RCT	Randomised controlled trial
UK	United Kingdom
USA	United States of America

Abstract

New health technologies seeking National Health Service funding in England are subject to rigorous evaluation of cost-effectiveness using established economic evaluation methods. Changes to the organisation and delivery of health services, including changes to health policy, are funded from the same budget as these health technologies, yet undergo no such mandatory cost-effectiveness assessment. This has resulted in a lack of methodological development and evidence on the cost-effectiveness of large-scale changes to the organisation and delivery of health services.

This thesis aims to contribute to the development of methods for the economic evaluation of changes to the organisation and delivery of health services. This aim is achieved through the consideration of two recent, high-profile examples: a regional pay-for-performance programme, and a national initiative to extend emergency hospital services at weekends.

Methods for both ex-ante and ex-post evaluation are developed and applied. The literature pertaining to the economic evaluation of the two example programmes examined is reviewed and critiqued. Estimates of the costs and benefits associated with the two programmes are provided.

A framework for assessing the cost-effectiveness of pay-for-performance programmes is developed, including a methodology to quantify the impact of programmes in terms of quality-adjusted life years in the absence of primary data collection on health-related quality of life. Issues around defining the relevant counterfactual for programme evaluations over the longer-term are explored, along with the potential for wider spillover effects. Survival analysis techniques commonly employed in clinical trials are used to improve treatment effect estimates associated with policy initiatives.

The regional pay-for-performance programme was found to have likely represented a cost-effective use of resources during the first 18 months of its operation. The programme was found to be associated with a health gain of 5,227 quality-adjusted life years, generated at a total cost to commissioners of £13m. The programme’s longer-term benefits are, however, uncertain.

The planned costs of extending emergency weekend hospital services are compared to the maximum potential health benefit attainable from this extension. The treatment effect associated with weekend admission to hospital is re-examined, and the possibility that earlier estimates suffer from bias is demonstrated, due to the restricted focus on only the admitted patient population.

The evaluation of the planned extension of weekend emergency hospital services shows that there is insufficient evidence to suggest that the programme would represent a cost-effective use of resources. The maximum potential health benefit attainable was estimated to be between 29,727 and 36,539 quality-adjusted life years, whilst the programme would cost between £1.07bn and £1.43bn. This exceeds the maximum the National Health Service should be willing to spend to achieve a health gain of this size by £339m to £831m.

Based on the two example interventions evaluated, and review of two established frameworks outlining the principles of cost-effectiveness analysis, the principal challenges faced when conducting economic evaluations of changes to the organisation and delivery of health services are identified and discussed. The principal challenges are identified as: undertaking ex-ante evaluation; modelling the counterfactual and estimating the treatment effect; evaluating the impact in terms of quality-adjusted life years; assessing costs and opportunity costs; accounting for spillovers; and generalisability.

Changes to the organisation and delivery of health services should undergo rigorous cost-effectiveness evaluation, as is now mandatory for new health technologies. This thesis contributes to the development of methods to facilitate such evaluation.

Declaration

No portion of the work referred to in this thesis has been submitted in support of an application for another degree or qualification of this or any other university or other learning institution.

Copyright statement

Rationale for submitting the thesis in journal format

This thesis has been submitted in journal format, meaning that the six substantive chapters are presented in the style of journal publications (Chapters 2 – 7). It was agreed with the supervisory team that the areas of research examined in this thesis presented opportunities to publish individual research papers alongside completion of this thesis. The journal format was therefore felt to be the most appropriate.

Author contributions

To date, five of the six substantive chapters of this thesis have been published in peer-reviewed journals. The research presented in these published chapters was completed in collaboration with other individuals, including the supervisors of this thesis, and these papers are therefore co-authored. Details of authorship, my individual contribution, and citations for the published works are listed below for each chapter.

Chapters 1, 7, and 8 are sole authored.

Chapter 2 is published in *Health Economics.* I developed the idea for the paper, performed the systematic review and developed the analytical framework to guide the cost-effectiveness of pay-for-performance programmes. I conceived the discounted and quality-adjusted life expectancy (DANQALE) tariff methodology, and calculated the relevant tariffs. Søren Rud Kristensen undertook the empirical difference-in-differences analysis of mortality, length of stay and readmissions, and repeated this analysis for the tariffed variables. I performed the costing exercise. I wrote the first draft of the manuscript. All authors contributed to the writing and editing of the paper. Citation: **Meacock R**, Kristensen S, Sutton M. (2014). The cost-effectiveness of using financial incentives to improve provider quality: a framework and application. *Health Economics,* 23, 1-13.

Chapter 3 is published in *New England Journal of Medicine.* All authors were involved in the conception of the paper. Søren Rud Kristensen undertook the empirical analysis in close consultation with myself, Alex Turner and Matt Sutton. Søren Rud Kristensen wrote the first draft of the manuscript. All authors contributed to the writing and editing of the paper. Citation: Kristensen S, **Meacock R**, Turner A, Boaden R, McDonald R, Rowland M, Sutton M. (2014). Long-term effect of hospital pay for performance on mortality in England. *New England Journal of Medicine,* 371(6), 540-548.

Chapter 4 is published in *Medical Decision Making.* Myself, Mark Harrison, and Matt Sutton were involved in the conception of the paper. I performed the empirical analysis, and wrote the first draft of the manuscript. All authors contributed to editing of the paper. Citation: **Meacock R**, Sutton M, Kristensen S, Harrison M. (2017). Using survival analysis to improve estimates of life year gains in policy evaluations. *Medical Decision Making,* 37, 415-426.

Chapter 5 is published in *Health Economics.* Myself and Matt Sutton conceived the idea for the paper. I reviewed the evidence on seven-day services, and estimated the costs of the policy. Matt Sutton estimated the potential benefit of introducing service extensions in England. I wrote the first draft of the manuscript. All authors contributed to editing of the paper. Citation: **Meacock R**, Doran T, Sutton M. (2015). What are the costs and benefits of providing comprehensive seven-day services for emergency hospital admissions? *Health Economics,* 24(8), 907-912.

Chapter 6 is published in *Journal of Health Services Research and Policy.* Myself and Matt Sutton conceived the idea for the paper. Matt Sutton performed the analysis whilst I provided substantial input regarding the analytical approach taken. I wrote the first draft of the manuscript. All authors contributed to editing of the paper. Citation: **Meacock R,** Anselmi L, Kristensen S, Doran T, Sutton M. (2017). Higher mortality rates amongst emergency patients admitted to hospital at weekends reflect a lower probability of admission. *Journal of Health Services Research and Policy*, 22(1), 12-19.

1. Introduction

1.1 Economic evaluation

Economic evaluation provides a framework to assess both the costs and consequences of alternative healthcare programmes (Drummond et al., 2015). It requires systematic identification, measurement, and valuation of the inputs and outcomes of two or more alternative courses of action, and the subsequent comparative analysis of these. The purpose of economic evaluation is to inform decisions regarding the optimal allocation of scarce healthcare resources. The benefits of a programme are compared to the opportunity costs of the resources required. These opportunity costs are the value of the benefits foregone elsewhere in the system as a result of committing resources to the programme in question.

Whilst the identification of costs is similar across economic evaluations, there are different approaches to the quantification of the consequences of alternative programmes. Cost-utility analysis (CUA) is now the most widely published form of economic evaluation (Drummond et al., 2015). CUA involves quantifying the benefits of a programme in terms of quality-adjusted life years (QALYs), which represent a generic measure of health. Assessing benefits in this common metric allows comparisons to be made across interventions in different areas of healthcare covering a variety of patient groups. CUAs take an extra-welfarist perspective, where the agreed objective is to maximise health, as measured using QALYs, subject to the budget constraint (Frew, 2017; Griffin et al., 2009).

Decision-making bodies in countries across Europe, in addition to Australia and Canada, have formally incorporated CUA into their health technology assessment (HTA) processes to inform both coverage and reimbursement decisions (Sanders et al., 2016). The most influential of these agencies has been the National Institute for Health and Care Excellence (NICE), who provide national guidance and advice to improve health and social care in England (NICE, 2017). Health technologies are subject to rigorous evaluations of cost-effectiveness using standard methodology set out in the NICE reference case before funding recommendations are made (NICE, 2013). Clinical commissioning groups, NHS England, and local health authorities are required by law to comply with NICE technology appraisal guidance that recommends a health technology is made available (NICE, 2014). This extensive appraisal process is intended to introduce an explicit trade-off between costs and benefits into the decision-making process, ensuring that the health technologies funded represent value for money.

1.2 The current scope of economic evaluation in England

In England, the NICE technology appraisal process covers medicinal products, medical devices, diagnostic techniques, surgical procedures or other therapeutic techniques, therapeutic technologies other than medicinal products, systems of care, and screening tools (NICE, 2014). This thesis is concerned with changes to the organisation and delivery of health services, including changes to health policy, which are not covered by these appraisal guidelines. I will refer to these as ‘service interventions’ for ease of reading. Service interventions are funded from the same National Health Service (NHS) budget as health technologies, yet are not required to undergo a mandatory cost-effectiveness assessment.

Although the principles of economic evaluation are applicable to all areas of healthcare spending, in practice, developments in this field have focused on HTA of pharmaceutical products (Drummond et al., 2009; Griffin et al., 2009; Tunis et al., 2003). The focus on pharmaceuticals may be the result of the greater availability of research funding in this area (Franken et al., 2012), political influence (Franken et al., 2012), an attempt to protect against the profit-maximising motives of pharmaceutical companies (Ciani et al., 2017; Franken et al., 2012), or have occurred simply because larger scale service changes represent a significantly more complex challenge to evaluators. Irrespective of the cause, the differing levels of scrutiny applied to spending on service interventions as opposed to health technologies is likely to result in allocative inefficiency in the healthcare system.

Recognition of the need to evaluate the impact of service interventions has grown over recent years, and a number of useful guidelines have been produced (Craig et al., 2008; Lamont et al., 2016; Medical Research Council, 2009, 2008; Raine et al., 2016). These provide clear and authoritative summaries of the range of methods available to evaluate the effects of changes to health services. However, none of these guidelines extend to economic evaluation.

1.3 Aims and objectives

The overall aim of this thesis is to contribute to the development of methods for the economic evaluation of changes to the organisation and delivery of health services.

This aim is achieved through the consideration of two recent, high-profile example service interventions: a pay-for-performance programme; and seven-day hospital services. With these examples in mind, I pursue the following objectives:

a) Review and critique the existing literature pertaining to the economic evaluation of the two example service interventions studied
b) Motivated by the two example interventions, develop methods for the economic evaluation of service interventions
c) Produce evidence on the costs and benefits associated with two example service interventions studied
d) Identify the principal challenges faced when conducting economic evaluations of service interventions

1.4 Example service interventions studied in this thesis

The first service intervention examined is the Advancing Quality (AQ) pay-for-performance (P4P) programme (Sutton et al., 2012). This initiative introduced financial incentives for hospitals to improve the delivery of health care. It was implemented in October 2008 with the objective of improving the quality of care provided by hospitals for five specific health conditions. AQ was introduced in the North-West region of England only, and is examined ex-post. The quasi-experimental nature of the programme's introduction, its focus on a small number of specific patient groups, and the ex-post approach taken makes this a useful foundation upon which to begin the methodological developments.

The second service intervention examined is the planned phased introduction of seven-day services for emergency hospital care in England (NHS England, 2017). This initiative aims to ensure access to the same high quality of care for patients on every day of the week, with a specific focus on improving the quality of care provided at weekends (NHS England, 2016). This is in response to findings of a 'weekend effect', where patients admitted to hospital in an emergency at the weekend have been found to experience higher mortality rates than those admitted during the week (NHS England, Seven Days a Week Forum, 2013a). The seven-day services programme involves a reorganisation of the way emergency hospital care is delivered, with an increase in service provision at weekends. Whilst there are separate initiatives to extend out-of-hours primary care provision, I focus solely on the emergency hospital care aspect of the policy. This is a national policy covering all patients admitted to hospital in an emergency at weekends, and is examined ex-ante, increasing the complexity of the evaluation and therefore making this a suitable example upon which to further methodological developments.

1.5 Overview of the thesis

This thesis consists of six substantive chapters, in addition to this introduction chapter, and a final concluding discussion chapter. As the thesis has been prepared in journal format, the six substantive chapters are presented in the style of journal publications, with their own introduction and discussion sections. Each of the six substantive chapters examine elements necessary for the economic evaluation of changes to the organisation and delivery of health services, with reference to the two service interventions selected for investigation. Each chapter is therefore focused on examining the costs and/or effects of a change to the organisation and delivery of health services.

Chapter 2 begins by systematically reviewing the literature on the economic evaluation of P4P programmes. An analytical framework is developed to guide the assessment of the cost-effectiveness of P4P initiatives, highlighting the issues that should be considered when undertaking such evaluations. Within this, a method is proposed to quantify the effects of service interventions in terms of QALYs in the absence of primary data collection on health-related quality of life. This framework is then applied to the AQ scheme to evaluate its cost-effectiveness during the first 18 months of operation.

Chapter 3 considers the longer-term impact of AQ on patient mortality over a further 24 months. Issues around defining the relevant counterfactual over the longer-term period are examined, as the programme structure changed and there was potential for quality improvement efforts to spillover onto wider patient groups.

Chapter 4 then revisits the methods used to quantify the effects of service interventions on QALYs in Chapter 2, further developing this approach. QALYs are comprised of two components covering both the quality and length of life (Drummond et al., 2015; Gold, 1996). Chapter 4 focuses on the length of life component, demonstrating how survival analysis techniques commonly employed in clinical trials can be used to improve treatment effect estimates associated with policy initiatives. These methods are applied to the AQ scheme to improve upon the previous estimates of the impact of the programme on survival.

Chapter 5 progresses to ex-ante evaluation, applied to the seven-day services policy. The evidence underpinning the policy is first examined, and estimates of its potential benefits produced. These are compared to the estimated planned costs of introducing seven-day services to evaluate whether the policy would likely be deemed to represent a cost-effective use of resources if it were subject to the same assessment rules as those applied in NICE technology appraisals.

Chapter 6 provides a critical assessment of the approach taken to estimate the weekend effect in previous research used to support the seven-day services policy. The possibility is explored that earlier estimates of the weekend effect are biased by the focus only on admitted patients. The scope of the analysis is expanded to include all patients attending accident and emergency (A&E) departments. The impact of variations in admission volumes on mortality for weekend admissions is investigated.

Chapter 7 identifies the principal challenges faced when conducting economic evaluations of service interventions. The specific challenges faced in the preceding chapters of this thesis are reviewed alongside two established frameworks outlining the principles of cost-effectiveness analysis. The principal challenges are discussed, and recommendations for overcoming them provided.

Chapter 8 provides the final discussion. The key findings of the thesis are summarised, and the strengths and weaknesses discussed. I also present the implications of the work and suggest directions for future research in this area.

2. The cost-effectiveness of using financial incentives to improve provider quality: A framework and application

Abstract

Despite growing adoption of pay-for-performance programmes in health care, there is remarkably little evidence on the cost-effectiveness of such schemes. We review the limited number of previous studies and critique the frameworks adopted and the narrow range of costs and outcomes considered, before proposing a new more comprehensive framework, which we apply to the first pay-for-performance scheme introduced for hospitals in England. We emphasise that evaluations of cost-effectiveness need to consider who the residual claimant is on any cost savings, the possibility of positive and negative spillovers, and whether performance improvement is a transitory or investment activity. Our application to the Advancing Quality initiative demonstrates that the incentive payments represented less than half of the £13m total programme costs. By generating approximately 5,200 quality-adjusted life years and £4.4m of savings in reduced length of stay, we find that the programme was a cost-effective use of resources in its first 18 months.

This chapter is published in *Health Economics.*

Citation: Meacock R, Kristensen S, Sutton M. (2014). The cost-effectiveness of using financial incentives to improve provider quality: a framework and application. *Health Economics,* 23, 1-13.

2.1 Introduction

Pay-for-performance (P4P) schemes, which link financial payments by purchasers to the quality of care supplied by health care providers, have grown in popularity over recent years. Care quality is commonly measured using pre-specified performance measures, which are often clinical processes judged to represent best practice or, less frequently, measures of outcome. Where clinical process measures are used, it is hoped that this will produce superior health outcomes for patients. Improving quality and outcomes may also reduce future health care costs. Despite much research by economists on this topic, there remains remarkably little evidence on the cost-effectiveness of such schemes.

A recent commentary highlights the ‘curious’ focus of research to date on the effectiveness of P4P schemes, with a neglect of their costs, and therefore cost-effectiveness (Maynard, 2012). This gap in the evidence base is also noted by a number of reviews. Greene and Nash (2009) provide an overview of the literature on P4P published between 2004 and 2008. Of the 100 articles included in their annotated bibliography, only three are grouped under the heading of ‘cost analysis’ (Curtin et al., 2006; Nahra et al., 2006; Parke, 2007). Mehrotra and colleagues (2009) systematically reviewed the evidence on hospital-based P4P programmes, stating there to be approximately 40 such schemes targeted at inpatient care. Despite this, only eight formal evaluations were found, covering just three different schemes. Of these eight published studies, just one (Nahra et al., 2006) attempted to estimate cost-effectiveness.

Most recently, Emmert et al. (2011) presented a systematic review of economic evaluations of P4P, critically assessing the identified studies on their methodological quality according to the widely used Drummond and Jefferson (1996) checklist. They identify nine studies, of which only three were categorized as full economic evaluations (An et al., 2008; Kouides et al., 1998; Nahra et al., 2006), and six as partial evaluations (Curtin et al., 2006; Lee et al., 2010; Norton, 1992; Parke, 2007; Rosenthal et al., 2009; Ryan, 2009). Of these six, four were deemed to be partial evaluations as they examined both the costs and effects of the P4P programmes under consideration but failed to make an explicit link between the two (Lee et al., 2010; Norton, 1992; Rosenthal et al., 2009; Ryan, 2009). The remaining two partial evaluations were simple cost comparisons, examining only the financial implications of the schemes in question (Curtin et al., 2006; Parke, 2007). This comprehensive review concluded that, on the whole, studies to date are methodologically flawed, failing to incorporate the full range of costs and consequences relevant to the evaluation of P4P.

Concerns regarding the value for money of the Quality and Outcomes Framework (QOF) in the UK led the Department of Health to commission a report in which a conceptual framework was developed to assess the cost-effectiveness of QOF indicators (Mason et al., 2008; Walker et al., 2010). This framework takes account of the cost of providing the incentivised interventions along with the incentive payments and the value of the health benefits achieved, but fails to incorporate the administrative costs associated with running the scheme. It also only considers the direct costs and benefits of changes in the incentivised measures and does not account for other changes in provider behaviour. Finally, it simulates the effects of better performance on the incentivised measures using published estimates of average effects and therefore does not reflect incremental changes. Whilst it is fundamentally important to ensure that the treatments incentivised by P4P programmes are themselves cost-effective, even after the additional cost of the incentive payments are considered, we believe it is necessary to take this a step further and consider whether P4P programmes as a whole represent a cost-effective use of resources.

Therefore, we aim to develop an analytical framework to guide the assessment of the cost-effectiveness of P4P programmes, highlighting the issues that should be considered when undertaking such evaluations. We first critique the narrow range of costs and effects considered by studies to date. This leads us to propose a new more comprehensive framework, highlighting the various cost categories that should be considered beyond the incentive payments themselves, along with issues such as who the residual claimant on any cost savings may be. Finally, we apply this framework to the first P4P scheme introduced for hospitals in the UK, the Advancing Quality (AQ) programme. The introduction of this scheme has been shown to have been associated with a significant reduction in mortality in the short-term (Sutton et al., 2012). We use our framework to show what additional analyses are required to assess whether the scheme was cost-effective. In particular, we consider how to convert the mortality reductions to gains in quality-adjusted life years (QALYs), what direct set-up and running costs to include, and estimate other indirect impacts on health service costs.

2.2 Methods

2.2.1 Critiquing previous evaluations

The recently published Emmert et al. (2011) review systematically appraised the quality of the economic evaluation literature. We present a brief commentary on the lack of methodological consistency between studies, focusing on the narrow range of costs and effects considered. Studies were identified from the previously mentioned review, and the search strategy used was run again in September 2012 to ensure that no new articles were missed. Studies known to the authors but not included in the Emmert et al. review were also included if they assessed the costs of P4P schemes. Details were extracted regarding the setting of the evaluation, the perspective taken, the main cost categories included and omitted, and the outcomes examined.

2.2.2 Developing the analytical framework

This appraisal of previously published evaluations was then used to develop a more comprehensive framework for assessing the cost-effectiveness of P4P schemes. The methodological issues brought to light in this first section were combined with the standard principles of cost-effectiveness analysis outlined in already established frameworks (e.g. Drummond, 2005; NICE, 2013) to provide a more specific framework to guide the evaluation of P4P programmes.

2.2.3 Applying the framework to the Advancing Quality initiative

We demonstrate the proposed framework by applying it to the AQ initiative that began in the North West of England in October 2008. We focus on the first 18 months, after which it was absorbed into the national Commissioning for Quality and Innovation (CQUIN) scheme (Department of Health, 2010). The programme aimed to improve quality in participating hospitals by paying for performance on 28 indicators across five health conditions. AQ ran in the North West of England only, and participation was universal within this region. We first discuss the issues raised in our framework in relation to the evaluation of AQ, before presenting estimates of the cost and effects of the programme.

We analyse mortality within 30 days of admission, emergency readmissions within 30 days, and length of stay (LOS) for three of the five incentivised conditions (acute myocardial infarction (AMI), heart failure and pneumonia). We exclude coronary artery bypass grafting (CABG) and hip and knee replacement as the mortality rate was below 2% for these procedures during the pre-intervention period. We use patient-level Hospital Episode Statistics data for patients admitted for one of these three AQ conditions in the period 1st April 2007 to 31st of March 2010, covering 18 months before and 18 months after the introduction of the programme. For the analysis of readmissions, we also include readmissions that occurred in April 2010. Our sample consists of 856,715 patients (662,458 patients for readmissions as we exclude patients not discharged alive) treated for one of the three conditions we examine at one of 154 hospitals across England[1]. Of these, 24 hospitals were in the North West of England, and thus subject to AQ, with the remaining 130 located in other regions of England, and therefore not subject to the policy. We evaluate the effects of AQ using a between-region difference-in-differences (DiD) analysis, comparing changes in outcomes in the North West to the changes in outcomes in the rest of England. The analysis was carried out at hospital level using weighted least squares on quarterly observations of risk-adjusted in-hospital mortality, readmission and mean LOS, allowing for hospital fixed effects and for time trends using quarterly dummy variables. The risk adjustment for each of the three outcomes of interest was conducted at patient level. The model for identifying changes in outcome after the introduction of AQ takes the form:

$$y_{jt} = \mathrm{a} + u_j + \nu_t + \delta D_j^1 \times D_t^2 + \varepsilon_{jt} \qquad (1)$$

with y_{jt} being the risk-adjusted outcome of interest at hospital j in quarter t, u_j the hospital fixed effects, ν_t the time fixed effects and ε_{jt} the residual term that is randomly distributed with a zero mean. The dummy variable D_j^1 equals one if the hospital is located in the North West, and zero otherwise. The variable D_t^2 equals one for all quarters after the introduction of AQ and zero beforehand. Our main interest is in the coefficient on the interaction of these two variables, δ. The main effects of D_j^1 and D_t^2 are not included, as they are perfectly collinear with the included time and hospital fixed effects.

[1] Hospital is used as shorthand for hospital Trust throughout this chapter

As well as considering changes in these variables in natural units, we repeat the DiD estimation using variables to which ‘tariffs’ have been applied. We apply a discounted and quality-adjusted life expectancy (DANQALE) tariff to the mortality outcome and the cost tariffs used in the national activity-based financing programme (‘Payment by Results’) to the readmissions and LOS.

The DANQALE tariff is stratified by single year of age (18-100 years) and sex. Sex-specific life expectancy estimates at each single year of age are taken from the 2008-10 Interim Life Tables from the Office for National Statistics (ONS) (2011). The age-sex specific quality of life adjustments are sourced from mean values of the EQ-5D index reported by respondents to the 2006 wave of the Health Survey for England. We calculate the DANQALE (Q_{ia}) for each individual *i* in each age-sex group *a* as:

$$Q_{ia} = (1 - m_i)\sum_{k=a}^{L_a} q_k (1 + r)^{-(k-a)} \qquad (2)$$

where m_i equals 1 if the individual dies within 30 days and 0 otherwise; *k* indexes ages from age *a* to the life expectancy of an individual currently aged *a* (L_a); q_k is health-related quality of life at age *k*; and *r* is the discount rate. We use an annual discount rate of 3.5% as specified by the National Institute for Health and Care Excellence (NICE) in their reference case (NICE, 2013). To cost LOS, we apply to each individual’s LOS the 2009/10 per-diem tariffs for days above the trim point for the main healthcare resource group (HRG) for which they were admitted. Readmissions are costed using the 2009/10 tariff prices for the main HRG for which the individual is admitted on re-admission.

A critical assumption of DiD is that the changes in the control hospitals are an appropriate counterfactual for the changes in the treated hospitals that would have occurred without the programme. We undertook pre-trends tests for all of the raw and tariffed outcomes and failed to reject the null-hypothesis of equal pre-trends at the 5% significance level for all conditions and outcomes bar LOS for heart failure patients (Table 1).

Table 1: Descriptive statistics for the North West and the Rest of England, by health condition and time period

Condition	**North West region**			**Rest of England**				
	Before introduction	After introduction	Difference	Before introduction	After introduction	Difference	Pre-trend tests	
AMI								
Patients, n	20,092	18,762	-1,330	104,912	101,479	-3,433		
Mortality rate, %	12.4	11.0	-1.4	11.0	10.7	-0.3	-0.4	[-1.02,0.20]
Readmission rate, %	11.9	12.1	0.2	10.9	11.1	0.2	-0.3	[-0.91,0.25]
Average LOS, days	9.3	8.5	-0.8	8.0	7.7	-0.3	-0.07	[-0.28,0.14]
Heart failure								
Patients, n	15,446	15,476	30	83,546	86,569	3,023		
Mortality rate, %	17.9	16.6	-1.3	16.6	16.1	-0.6	0.3	[-0.44,1.02]
Readmission rate, %	17.8	18.4	0.7	17.3	17.0	-0.2	0.0009	[-0.80,0.81]
Average LOS, days	11.9	11.2	-0.7	11.4	11.0	-0.5	-0.3	[-0.66,-0.04]
Pneumonia								
Patients, n	28,275	36,428	8,153	150,526	195,204	44,678		
Mortality rate, %	28.0	25.9	-2.2	27.2	26.3	-0.9	-0.1	[-0.72,0.46]
Readmission rate, %	15.1	15.7	0.6	13.2	13.7	0.5	-0.2	[-0.81,0.42]
Average LOS, days	12.8	11.8	-1.0	11.8	11.4	-0.4	-0.2	[-0.47,0.02]

Notes: AMI, acute myocardial infarction. LOS, length of stay.

The pre-trend tests are the estimated difference between the linear quarterly trends in the North West and rest of England. 95% confidence intervals in brackets

2.3 Results

2.3.1 Previous evaluations

We identified 14 studies examining the cost of P4P schemes (Table 2). The majority of these schemes operated in the United States of America (USA) [1-8,10-11][2], with two in the United Kingdom (UK) [13-14], and one in each of Germany [9] and China [12]. The most common setting for the programmes under evaluation were primary care clinics [2,4,6-9,12-14], followed by hospitals [3,5,11]. Nine of the 10 USA evaluations were undertaken from the perspective of the health plan [1-5,7-8,10-11], with one extending this to include the plan's enrolees [7] and another also considering the providers' perspective [1]. Just one evaluation was performed purely from the providers' perspective [9], and the remaining three from that of government-run health systems [12-14]. The range of costs included by many of the studies were however inconsistent with their stated perspectives, often failing to encompass all relevant cost categories. Just two evaluations clearly incorporated the costs associated with the development and set-up of the P4P schemes in question [4,8], and only six included the ongoing running costs [4-9]. Seven studies made some attempt to measure the increased costs associated with providing the incentivised treatments [1,4,6,8-9,12,14]. Five studies failed to consider any costs beyond the incentive payments themselves [2-3,10-11,13].

Of the 14 studies examining the cost of P4P programmes, 11 also made some attempt to estimate the effects of the schemes [1-3,5,8-14]. However, the range of effects considered was narrow. The incentivised performance measures were by far the most commonly used metrics of effect, with all but one evaluation reporting results on these process or clinical measures [11]. Four evaluations considered only these incentivised measures [2-3,8-9] and made no attempt to link quality improvements to health outcomes. Three studies examined intermediate outcomes, such as hospitalisations and LOS [1,10,12], and three examined mortality [1,5,11]. Just two of the evaluations attempted to express the effects of P4P schemes in terms of QALYs [5,14], and only one looked at the potential effects on non-incentivised areas of care [13].

The omission of relevant cost categories by many previously conducted evaluations, along with the lack of evidence regarding the effects on health outcomes, means that conclusions regarding the value for money of the programmes in question cannot be made.

[2] Numbers in [ ] refer to the study number given in Table 2 and are used to enable ease of reading for this summary

Table 2: Summary of previous economic evaluations of pay-for-performance schemes

Study no.	**1**	**2**	**3**	**4**	**5**	**6**	**7**
First author, year	Norton (1992)	Kouides (1998)	Kahn (2006)	Curtin (2006)	Nahra (2006)	Brown (2007)	Parke (2007)
Country	USA	USA	USA	USA	USA	USA	USA
Setting	Nursing homes	Primary care clinics	Hospitals	Primary care clinics	Hospitals	Primary care clinics	Primary care clinics
Perspective	Health plan & providers	Health plan	Health plan	Health plan	Health plan	Providers	Health plan (& its enrolees)
Costs included:							
Development/set-up costs	X	X	X	✓	X	?	X
Running costs	?	X	X	✓	✓	✓	✓
Treatment costs	✓	X	X	✓	X	✓	?
Incentive payments	✓	✓	✓	✓	✓	✓	✓
Outcomes:							
Incentivised measures	✓	✓	✓	X	✓	X	X
Intermediate	✓	X	X	X	X	X	X
Mortality	✓	X	X	X	✓	X	X
QALYS	X	X	X	X	✓	X	X
Non-incentivised care	X	X	X	X	X	X	X

Notes: X=not reported, ?=unclear/lack of detail, ✓=reported. QALYs, quality-adjusted life years.

Table 2 continued

Study no.	**8**	**9**	**10**	**11**	**12**	**13**	**14**
First author, year	An (2008)	Salize (2009)	Rosenthal (2009)	Ryan (2009)	Lee (2010)	Sutton (2010)	Walker (2010)
Country	USA	Germany	USA	USA	China	UK	UK
Setting	Primary care clinics	Primary care clinics	Prenatal care	Hospitals	Primary care clinics	Primary care clinics	Primary care clinics
Perspective	Health plan	Health plan	Health plan	Health plan	National health insurance system	National health insurance system	National health insurance system
Costs included:							
Development/set-up costs	✓	X	X	X	X	X	X
Running costs	✓	✓	X	X	X	X	X
Treatment costs	✓	✓	X	X	✓	X	✓
Incentive payments	✓	✓	✓	✓	✓	✓	✓
Outcomes:							
Incentivised measures	✓	✓	✓	X	✓	✓	✓
Intermediate	X	X	✓	X	✓	X	X
Mortality	X	X	X	✓	X	X	X
QALYS	X	X	X	X	X	X	✓
Non-incentivised care	X	X	X	X	X	✓	X

Notes: X=not reported, ?=unclear/lack of detail, ✓=reported. QALYs, quality-adjusted life years.

2.3.2 Analytical framework

2.3.2.1 Perspective

The relevant perspective will depend upon the institutional arrangements into which the P4P programme is introduced. In the UK this is likely to be that of the National Health Service (NHS) and personal social services (PSS), consistent with the perspective specified by NICE in their reference case (NICE, 2013). The perspective should be clearly stated and the range of costs and effects considered should be consistent with this.

As health care providers act as agents to payers (who can be thought of as principals under Principal-Agent Theory), and these payers in turn act as agents to customers/taxpayers, it is worth at least considering the perspective of providers as well as payers when determining the cost-effectiveness of P4P. It may therefore be relevant to consider not only whether it is cost-effective for the payer to run a P4P programme, but also whether it is cost-effective for providers to participate in and perform the tasks necessary to improve performance on the stipulated quality measures. Providers may incur substantial costs as a result of participating in P4P programmes, both in terms of the capital investments necessary to permit activities such as data collection and the cost of providing the incentivised treatment itself. Whilst some/all of these costs may be offset by the incentive payments, there is no guarantee that providers will actually receive bonuses as these are conditional upon performance. In some cases such schemes operate as a tournament, with only the top performers receiving a bonus payment, and under some programmes there may even be the possibility of financial sanctions if performance benchmarks are not met.

2.3.2.2 Comparator

A clear comparator is essential for any economic evaluation (Drummond et al., 2015), representing what would have happened in the absence of the programme. The relevant counterfactual will again depend upon the institutional arrangements. An important consideration is whether to compare to the same additional resources but paid in a different manner or whether to compare to no additional payments. This depends on whether we are interested in P4P as a payment mechanism or as a form of potential additional funding.

Ideally, the programme would first be introduced under conditions of randomisation, with providers being allocated to an intervention group receiving P4P or a control group. This would allow selection bias and confounding factors to be avoided. In practice however, P4P schemes are rarely launched in this way (Scott et al., 2011). It may be possible to employ a quasi-experimental design using providers not participating in the scheme as a comparator group if, for example, P4P has only been implemented in certain geographical areas. It is vital, however, that the analysis takes into account any potential sources of bias such as differing provider or patient characteristics between the groups. Alternatively, providers may be used as their own controls in a before-after study design, with observed outcomes before the implementation of P4P being projected forward in order to predict outcomes in the absence of the programme. Again, attempts must be made to control for potential sources of bias such as general time trends which may have also affected the outcomes under examination.

2.3.2.3 Cost categories

Whilst the incentive payments themselves are by far the most obvious cost component of P4P programmes, there are many other costs involved in design and implementation. Whilst their relevance and magnitude will differ between programmes, the following cost categories should be considered:

- Set up/development costs – e.g. staff time, infrastructure investment. These costs can be spread across the expected lifetime of the policy if this is known.
- Running costs – e.g. administration.
- Incentive payments
- Costs to providers of participating in the scheme – e.g. staff time, pharmaceuticals. The perspective of the evaluation will dictate whether these costs are relevant.
- Cost savings – e.g. reduced complications, LOS, readmissions. It is assumed that improving the quality of care will produce superior health outcomes, which in turn has consequences for future health care costs. These cost savings may fall on providers or commissioners depending on the payment rules, so it is important to consider who the residual claimant may be.

The above cost categories and examples are not exhaustive, and illustrate the many possible financial implications of P4P schemes beyond the incentive payments themselves. As with any economic evaluation, the likely magnitude of each cost category must be weighed up against the resources involved in accurate estimation. There may be justification for excluding certain costs if it is clear that either they are insignificant in comparison to the overall cost of the policy, or that their inclusion will simply further confirm the current conclusions, but this should nevertheless be discussed.

2.3.2.4 Opportunity cost

As with any economic evaluation, we are concerned with the opportunity cost of the resources used by a programme, which in the case of health care spending represents the possible health gains foregone through not providing alternative treatments. P4P programmes are not always financed by additional funds, but may instead involve a reallocation of current budgets or resources. For example, a percentage of the existing budget may be top-sliced to fund the incentive scheme, or the duties of existing members of staff may be changed to focus on the areas of care incentivised. Whilst this does not involve any additional spending, these resources still have an opportunity cost in terms of care displaced.

2.3.2.5 Outcomes

The main outcomes recorded for P4P programmes are the targeted quality measures upon which performance is judged. If these are process rather than outcome measures, then evidence on their link with health outcomes should be presented. Ideally, benefits would be expressed in terms of QALYs in order to permit comparison with standard cost-effectiveness thresholds (NICE, 2013; Walker et al., 2010).

However, since quality is multi-dimensional, the outcomes influenced by P4P programmes are likely to stretch beyond those captured by the targeted performance measures, with the potential for both positive and negative spillovers into non-incentivised areas of care. If incentives divert the existing efforts of providers away from non-incentivised areas of care rather than promoting additional effort in the targeted areas, this could result in unintended consequences for patients (Kelman and Friedman, 2009). Depending on how well the chosen performance indicators capture the desired outcomes, the provider’s degree of altruism, and to what extent effort on the incentivised and non-incentivised dimensions are substitutes or complements to the agent, it may even be undesirable to pay for performance (Holmstrom and Milgrom, 1991; Kaarboe and Siciliani, 2011). Gaming is also a possibility, where providers merely make their performance on the incentivised measures appear better than it actually is, normally through manipulation of the reporting systems used to record such performance. A broad range of outcomes extending beyond the incentivised measures should therefore be considered when evaluating P4P schemes in order to fully capture their effects, both intended and unintended.

2.3.2.6 Time horizon

As with any economic evaluation it is important to capture all of the relevant costs and consequences attributable to a programme, which are likely to span over a number of years. An interesting point to consider is the expected lifetime of P4P schemes, which are seldom stated, and their ability to induce continued quality improvements year-on-year. Whilst we may expect to observe performance improvements when P4P is first introduced, these may reach a ceiling after which little or no further quality improvements are achieved. It may then be relevant to consider the consequences of removing the financial incentives currently in place if they are failing to induce additional benefits. The effect of this removal will depend upon whether quality improvement is a transitory or investment activity. Quality could fall, perhaps even to levels below those observed before the introduction of P4P (Lester et al., 2010). Alternatively, incentivised behaviours may have become routine and therefore continue even after payments are withdrawn. If some of the benefits are sustained beyond the period of cessation of the incentive payments, the cost-effectiveness of the scheme will be underestimated by a restricted evaluation period.

2.3.3 Application to Advancing Quality

2.3.3.1 Perspective

We examine the cost-effectiveness of AQ from the perspective of the NHS, estimating the costs incurred by commissioners and the resulting health benefits achieved. However, we note that as the programme ran under a tournament system, only half of the hospitals received bonus payments at each payout. Providers may have incurred substantial participation costs yet received no financial rewards for their efforts.

2.3.3.2 Comparator

We take advantage of the fact that AQ was introduced through universal participation and in the North West of England only to employ a quasi-experimental design in which the rest of England acts as the comparator.

2.3.3.3 Cost categories

We seek to include all of the relevant costs incurred by commissioners. These include the one-off lump-sum grants given to providers to cover the investments in infrastructure necessary to enable the required data collection. As AQ was merged with another national P4P policy 18 months after its introduction, and the expected life time of this new policy is unknown, the entire set-up costs are attributed to the first 18 months. Thus, the estimate of the costs of AQ over the first 18 months represents the upper bound of the actual costs applicable to this period.

We also include the financial incentives paid out to providers, the ongoing running costs and other one-off costs incurred within the period. The general running costs include the contract with Premier Inc. who oversaw the scheme, the central AQ team, auditing activities, quality improving consultancies and other administrative costs. One-off costs include legal fees and other procurements. We examine the potential cost savings resulting from reduced readmissions and LOS, and discuss who the residual claimants on these savings are likely to have been.

2.3.3.4 Opportunity cost

AQ was financed by a reallocation of the North West commissioning budget and so did not result in any additional spending by payers. We cannot determine what this money would have been spent on in the absence of the policy, and so use the lower bound of the standard UK cost-effectiveness threshold of £20,000 per QALY to reflect opportunity costs.

2.3.3.5 Outcomes

Hospital performance on the incentivised process and clinical measures is reported annually on the AQ website (http://www.advancingqualitynw.nhs.uk). We examine whether there is evidence that adoption of the scheme has translated into better health outcomes for patients. We evaluate the effect of AQ for only three of the five conditions. This means that our estimates of the effects of AQ are conservative, representing the lower bound of the actual effects as they do not take into account any benefits achieved in the remaining two clinical areas. Our cost estimates, however, do include the costs of the AQ programme as a whole, as it was not possible to separate out the costs applicable to each clinical area. The resulting estimates of cost-effectiveness are therefore also conservative, and at this stage assume that no health benefits were attained for hip and knee replacement and CABG patients.

2.3.3.6 Time horizon

We analyse the costs over the 18 month period from October 2008 to April 2010 and discount the effects over the expected patient lifetimes using an annual discount rate of 3.5%.

2.3.4 Costs of Advancing Quality

The total cost of the programme to commissioners was just over £13million, with only £5million of this consisting of the financial incentives (Table 3). The ongoing running costs of £7million exceed the bonus payments, and make up the majority of the costs. This result reinforces the importance of considering the costs of P4P beyond the incentive payments themselves. If, like five of the 14 previous studies identified in our earlier critique, we had failed to include any costs other than the bonuses paid out to the top performing hospitals, we would have underestimated the true cost of AQ by over 60%. Even if we exclude the set up costs, which it could be argued should be spread across a number of years, the incentive payments themselves still only represent 42% of the cost of the programme.

Table 3: Costs of the Advancing Quality programme during its first 18 months of operation

Activity	**Costs**
Set up costs	£990,000
Incentive payments	£5,054,489
Programme running costs	£7,015,531
One-off programme costs	£9,844
Total costs	**£13,069,864**

2.3.5 Effects of Advancing Quality

We estimate a statistically significant reduction in mortality and LOS associated with the introduction of AQ (Table 4), which is statistically significant for pneumonia only when the three conditions are analysed individually. Readmission rates are unchanged. There are also statistically significant reductions in DANQALE and cost-tariffed LOS (Table 5), again statistically significant for pneumonia only when the three conditions are analysed separately.

Table 4: Difference-in-differences estimates of Advancing Quality on percentage risk of mortality, readmission and days of hospital stay

	Mortality	Readmissions	LOS
Total incentivised	-0.9*** [-1.4,-0.4]	0.2 [-0.3,0.7]	-0.3** [-0.5,-0.1]
AMI	-0.3 [-1.0,0.4]	0.1 [-0.7,0.9]	-0.3 [-0.6,0.1]
Heart failure	-0.3 [-1.2,0.6]	0.7 [-0.4,1.9]	-0.2 [-0.6,0.3]
Pneumonia	-1.6*** [-2.4,-0.8]	-0.0 [-0.8,0.7]	-0.5** [-0.8,-0.1]

Notes: Between region difference-in-differences estimates

95% confidence intervals in brackets

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

LOS, length of stay. AMI, acute myocardial infarction.

Table 5:Difference-in-differences estimates of Advancing Quality on QALY-tariffed mortality and cost-tariffed readmissions and days of hospital stay

	Discounted QALYs	Readmissions £	LOS £
Total incentivised	0.07*** [0.04,0.11]	9.0 [-5.2,23.3]	-62.4** [-102.4,-22.3]
AMI	0.04 [-0.01,0.10]	11.2 [-9.0,31.4]	-58.1 [-118.1,2.0]
Heart failure	0.00 [-0.06,0.07]	21.8 [-14.8,58.3]	-31.6 [-111.7,48.4]
Pneumonia	0.13*** [0.06,0.19]	0.7 [-20.2,21.6]	-82.1* [-146.0,-18.2]

Notes: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

QALY, quality-adjusted life years. LOS, length of stay. AMI, acute myocardial infarction. LOS costed at per diem HRG tariff. Readmissions costed at the HRG tariff of the readmission. QALYs estimated on the basis of age- and sex-specific life expectancy and health-related quality of life estimates for age at admission.

2.3.6 Cost-effectiveness of Advancing Quality

We estimate a reduction of 649 deaths[3] and a gain of 5,227 QALYs as a result of the programme (Table 6). At a QALY value of £20,000, this equals an estimated health gain worth £105 million.

Our estimates suggest that AQ resulted in 22,802 fewer days in hospital, saving £4.4million. Due to the structure of the payment system in operation in England, where payment for a hospital admission is largely independent of LOS, these cost savings would be claimed mostly by providers rather than commissioners. For readmissions, we estimate a statistically insignificant £0.6 million increase in costs across all conditions.

Table 6: Total effects of Advancing Quality on outcomes in raw and tariffed units

	Total outcome changes in natural units			**Total benefits/costs**		
Condition	Mortality (deaths)	Readmissions	LOS (days)	∆QALYs	Readmissions £m	LOS £m
Total incentivised	-649	996	-22,802	5,227	0.6	-4.4
AMI	-60	168	-4,787	778	0.2	-1.1
Heart failure	-44	644	-2,493	26	0.3	-0.5
Pneumonia	-580	-47	-16,540	4,701	0.0	-3.0

Notes: LOS, length of stay. QALYs, quality-adjusted life years. AMI, acute myocardial infarction.

LOS costed at per diem HRG tariff. Readmissions costed at the HRG tariff of the readmission. QALYs estimated on the basis of age- and sex-specific life expectancy and health-related quality of life estimates for age at admission.

[3] This figure equals that which would be produced by Sutton et al.'s (2012) between-region DiD estimation but is fewer than the 890 deaths arising from their triple-difference models.

2.4 Discussion

P4P schemes are increasingly being used by purchasers as a means to encourage providers to improve their quality of care. Research to date has focused on whether such programmes induce changes on the targeted quality measures, commonly neglecting the more pertinent issue of their effect on health outcomes and costs. After critiquing the narrow range of costs and effects considered by previous evaluations, we developed an analytical framework to guide the future assessment of the cost-effectiveness of P4P programmes.

Our application of this framework to the AQ initiative reinforces the importance of considering costs beyond the incentive payments themselves, as failing to do so would have led us to include only 40% of the costs of the scheme from the commissioners' perspective. We have also estimated the incremental effects of AQ on mortality, readmissions and LOS directly, rather than relying on simulation modelling of the scheme's consequences. We observed statistically significant reductions in mortality and LOS attributable to the programme, and converted the mortality reductions into expected QALY gains. Despite incorporating a wide range of programme costs into our evaluation, we still find it likely that AQ represented a cost-effective use of resources during the 18 month period under examination at standard UK threshold values. Crude estimates put the monetary value of the estimated QALYs gained at £105million, far exceeding the £13million spent by commissioners on the programme.

Some biases may be present in our analysis for a number of reasons, most of which lead us to underestimate rather than overestimate the benefits attributable to the AQ programme. Firstly, we were only able to estimate outcomes for three of the five incentivised conditions and therefore make the conservative assumption that no benefits are produced for hip and knee replacement or CABG patients. Secondly, we are unable to estimate any 'pure' quality of life effects not associated with mortality. Thirdly, we assume that any observed improvements in quality of care are transitory and will not affect future patients. However, the use of age-sex specific DANQALE estimates from the general population is likely to over-estimate the health gains enjoyed by the additional survivors since the average life expectancy and health-related quality of life of individuals admitted to hospital for AMI, heart failure and pneumonia are likely to be lower than that of the general population. Nonetheless, just one QALY on average would need to be produced as a result of each death averted for AQ to be deemed cost-effective at the standard threshold of £20,000 for the value of a QALY.

We also estimated cost savings of £4.4million as a result of reduced LOS. Due to the structure of the payment system in operation, these cost savings would have been accrued largely to providers rather than payers. It is therefore rather puzzling that providers required financial incentives from purchasers to encourage such quality improving behaviour, when this behaviour is likely to have reduced their own costs. One possible explanation is that providers required the additional technological information on what represents best practice to realise such savings. Alternatively, the cost of providing the improved care may outweigh the reduced LOS cost savings, and so in the absence of the financial incentives it may not be efficient for providers to engage in quality improving behaviour.

Although it appears that AQ is likely to have represented a cost-effective use of resources during the 18-month period we evaluated, an important consideration for policy makers is its ability to continue generating improvements in the long run. This concern applies to all P4P schemes. It may be that P4P should be seen as a vehicle to kick start quality improving behaviours in the short term, which will then become engrained into routine. Alternatively, the observed improvements may simply represent transitory effort increases which will fall away once the financial incentives are removed.

This is one of several aspects of P4P schemes about which there is little good quality evidence. These include: whether the incentives should be bonuses or fines; what size of incentive is required; whether payments should be made for outcomes or activities likely to lead to better outcomes; whether schemes should be tournaments or potentially reward all providers; and whether payment schedules should be linear or ‘stepped’.

The intended and unintended behavioural responses of providers have formed the main focus of most research on P4P, not whether it is cost-effective. Yet resources spent on P4P also have opportunity costs. There are several P4P schemes in the health sector that would be worthy of cost-effectiveness analysis. We hope that the framework we have proposed will be developed further and applied to these schemes in the future.

3. Long-term effect of hospital pay-for-performance on mortality in England

Abstract

Background: A pay-for-performance programme based on the Hospital Quality Incentive Demonstration was introduced to all hospitals in the North West region of England in 2008 and was associated with a short-term (18 month) reduction in mortality. We analysed the long-term effects of this programme, called Advancing Quality.

Methods: We analysed 30-day in-hospital mortality among 1,825,518 hospital admissions for eight conditions, three of which were covered by the financial incentive programme. The hospitals studied included the 24 hospitals in the North West region that were participating in the programme and the 137 elsewhere in England that were not participating. We used difference-in-differences regression analysis to compare risk-adjusted mortality for an 18 month period before the programme was introduced with subsequent mortality in the short-term (the first 18 months of the programme) and the longer-term (the next 24 months).

Results: Throughout the short-term and the long-term periods, the performance of hospitals in the incentive programme continued to improve and mortality for the three conditions covered by the programme continued to fall. However, the reduction in mortality among patients with these conditions was greater in the control hospitals (those not participating in the programme) than in the hospitals that were participating in the programme (by 0.7 percentage points; 95% confidence interval 0.3 to 1.2). By the end of the 42 month follow-up period, the reduced mortality in the participating hospitals was no longer significant (-0.1 percentage points; 95% confidence interval -0.6 to 0.3). From the short-term to the long-term, the mortality for conditions not covered by the programme fell more in the participating hospitals than in the control hospitals (by 1.2 percentage points; 95% confidence interval 0.4 to 2.0), raising the possibility of a positive spillover effect on care for conditions not covered by the programme.

Conclusions: Short-term reductions in mortality for conditions linked to financial incentives in hospitals participating in a pay-for-performance programme in England were not maintained.

This chapter is published in *New England Journal of Medicine.*

Due to the strict word count of the journal, some material was confined to a supplementary section for the published article. Where relevant, some of this supplementary material has been moved to the main body of the paper here in order to give a sufficient level of detail for this thesis.

Citation: Kristensen S, Meacock R, Turner A, Boaden R, McDonald R, Roland M, Sutton M. (2014). Long-term effect of hospital pay for performance on mortality in England. *New England Journal of Medicine,* 371(6), 540-548.

3.1 Introduction

Pay-for-performance (P4P) initiatives, which explicitly link financial incentives to the performance of health care providers, have been adopted in several countries in recent years (Eijkenaar, 2012; Paris et al., 2010). These programmes aim to improve the quality of care provided, which should result in better patient outcomes. However, evidence that improvements in health are realised in practice is currently lacking (Eijkenaar et al., 2013; Mehrotra et al., 2009; Rosenthal and Frank, 2006). Few programmes have been subjected to robust evaluation. Programmes that have been evaluated show modest and short-term improvements at best on measures of processes related to financial incentives (Grossbart, 2006; Lindenauer et al., 2007). There is particular concern about the long-term effects of P4P initiatives, since initial improvements in measures of quality that are associated with incentives may not be sustained (Campbell et al., 2009).

The largest hospital P4P initiative to be implemented to date, the Premier Hospital Quality Incentive Demonstration (HQID), failed to have a significant effect on mortality within either the first three (Glickman et al., 2007; Ryan, 2009) or six (Jha et al., 2012) years of its operation. Previous evaluations of the Advancing Quality (AQ) programme, an initiative based on HQID, showed that its adoption in one region of England in 2008 led to a clinically significant reduction in 30-day in-hospital mortality during the first 18 months (Sutton et al., 2012; Chapter 2). The fact that the programme had a more positive effect in England than in the United States has been attributed to the universal participation of hospitals within the region in England, the larger bonus payments, and the collaborative nature of the initiative, which led hospitals to make more general investments in quality improvement. We have extended this earlier analysis to consider the longer-term effects of the policy over an additional 24 months.

3.2 Methods

3.2.1 The incentive programme

Starting in October 2008, all 24 hospitals[4] providing emergency care in the North West region of England participated in the AQ programme. From the beginning, the initiative involved reporting on measures of quality of care related to clinical conditions in five categories: acute myocardial infarction (AMI), heart failure, pneumonia, coronary artery bypass grafting, and hip and knee surgery. Our analysis focused on the first three, which represent conditions for which patients are hospitalised on an emergency basis for treatment.

The financial reward system changed twice during the 3.5 year period under consideration. The first year was run as a pure tournament, with hospitals scoring in the top quartile on the quality metrics linked to incentives receiving a 4% bonus payment and those in the second quartile receiving a bonus of 2%. For the next 6 months, financial incentives were awarded on the basis of three criteria. Providers whose performance in this period was ranked above the median score

[4] Hospital is used as shorthand for hospital Trust throughout this chapter

from the first year were awarded an ‘attainment’ bonus. Those earning this attainment bonus were then eligible for two further payments, which were awarded to hospitals in the top quartile for improved performance and those in the top two quartiles for absolute performance. There were no penalties or withholding of a percentage of reimbursement for poor performers (those not qualifying for a performance payment) during these first 18 months.

After the first 18 month period, the structure of the financial incentives changed again. Instead of bonuses, a fixed proportion of the hospital’s expected income was withheld and paid out only if required performance thresholds were reached. The performance indicators remained the same, and required levels of achievement were based on the quality scores that had been achieved by each hospital in the first year of the AQ programme.

The total amounts of money potentially linked to performance were kept constant throughout the period. Bonuses of £3.2million were paid to hospitals in the North West region for the first year and bonuses of £1.6million were paid for the next six months. When the incentives changed from bonuses to penalties, the total potential losses for hospitals were £3.2million each year if all hospitals failed to meet all of the targets for the five AQ conditions.

The financial payments were made to the hospitals, not the clinical teams directly. Hospitals made some investments in additional staff in the specialties covered by the AQ programme when the existing staff made a convincing case that these investments were necessary, but these investments were not dependent on the receipt of bonuses.

The AQ programme included only hospitals in the North West region of England. Two other regions began to institute the AQ measures during the long-term period of our study, but they did not have access to any of the supporting mechanisms in place for the North West region. A small number of hospitals outside these regions also instituted a financial incentive programme for similar clinical conditions, but again with no supporting mechanisms. We included these hospitals in the control group in the main analysis and tested the effect of their exclusion in supplementary analysis (see section 3.2.3.5).

3.2.2 Data

Data on the quarterly performance of hospitals with respect to quality measures related to the incentive programme were obtained from the AQ administrators. Data on patient characteristics, co-existing conditions, and mortality were obtained from national Hospital Episode Statistics (Health and Social Care Information Centre, 2012). As in the short-term evaluation of AQ presented in Sutton et al. (2012), we used data for all patients in England who were admitted on an emergency basis for treatment of AMI, heart failure, or pneumonia. We obtained equivalent data for patients admitted in an emergency for one of five primary diagnoses that were not related to the incentives in the AQ programme or national programmes at any point during the study[5]: acute renal failure (International Classification of Disease (ICD)-10 codes beginning N17); alcoholic liver disease (K70); intracranial injury (S06); paralytic ileus and intestinal obstruction without hernia (K56); duodenal ulcer (K26). These conditions were chosen for the earlier Sutton et al. (2012) evaluation to meet the following criteria: no clinical linkage to any condition included in the programme; sufficient volume (more than 9,000 admissions in England per year); a 30-day mortality rate of more than 6%; and more than 80% of deaths within 30 days of admission occurring in hospital.

Data were obtained for patients admitted between 1st April 2007 and 31st March 2012. We divided the data into three periods: i) *Before* – the 18 months prior to the introduction of the scheme; ii) *Short-term* – the first 18 months of its operation; and iii) *Long-term* – months 19 to 42 of the programme. The final sample included 390,652 patients admitted for AMI, 338,921 for heart failure, 761,954 for pneumonia, and 333,991 for conditions not related to the incentives, treated at 161 hospitals across England. We analysed mortality occurring in any hospital within 30 days of admission.

During the data extraction process, we select only one admission record per patient during the last 30 days of their life. This ensures that we do not count deaths multiple times for patients with repeated hospital admissions during their last 30 days.

3.2.3 Statistical analysis

We calculated the expected risks of death using logistic regression models at the patient level that included sex and age, 31 coexisting conditions included in the Elixhauser algorithm, derived from secondary ICD-10 diagnostic codes (Quan et al., 2005), type of admission (emergency or transfer from another hospital), and the location from which the patient was admitted (own home or an institution). The analysis of risk-adjusted mortality was performed with data aggregated on the basis of the 3 month calendar period and the admitting hospital.

We used two analyses to test whether the incentives had an effect on mortality. The first was a between-region difference-in-differences (DiD) analysis that compared changes in mortality over time between the North West region and the rest of England for the incentivised and non-incentivised conditions. The second was a triple-difference analysis that compared the changes in mortality over time between the incentivised conditions in the North West region of England, then subtracted the changes in mortality over time between the North West and the rest of England for non-incentivised conditions.

We estimated the effects for the combination of all three incentivised conditions, for the combination of all five non-incentivised conditions, and for each incentivised condition individually.

[5] Pulmonary Embolism was included as a non-incentivised condition in the previous Sutton et al. (2012) study, but was not included in this analysis as it was incentivised in a separate national programme from April 2010

We weighted condition-specific mortality using total admissions over the entire study period to ensure that the combined mortality series did not reflect changes in the relative volumes of patients admitted for different conditions. To account for temporal trends, we included an indicator for each of the 20 quarter-year periods in all analyses. To account for differences among hospitals, we assigned an indicator to each hospital in the analyses of individual conditions and assigned an indicator to each combination of hospital and condition in the combined analyses. We estimated separate effects for the short-term and long-term periods by including terms for the interaction between the intervention group and each of the two post-implementation periods.

3.2.3.1 Risk-adjustment

We adjusted the observed mortality rates using expected mortality rates that were obtained from patient-level logistic regression models. For each health condition, the binary outcome at patient level was regressed on vectors of binary variables for sex and age-group interaction dummy variables (where age is measured in five-year bands), the first four digits of the full primary ICD-10 diagnosis code, 31 Elixhauser conditions, the admission method and the admission source. The predicted probabilities from these patient-level models were then averaged across strata defined by quarter of admission and admitting hospital.

For heart failure, we did not control for Elixhauser congestive heart failure codes as these diagnoses were included in the primary diagnoses. For AMI, we did not control for arrhythmia codes, as the incentive scheme was specifically designed to prevent these. In principle, we might have wished to control for pre-existing arrhythmias, but we could not distinguish between pre-existing arrhythmias and those which occurred while in hospital, so we excluded arrhythmias in the main analysis.

3.2.3.2 Regression methods

For each health condition group in the incentivised and non-incentivised conditions, separately and combined for the incentivised and non-incentivised conditions, we present three results. These measure the impact of the programme over three different time periods; i) the difference in mortality between the North West and the rest of England from before AQ was introduced to the short-term period, ii) from before AQ to the long-term period, and iii) the difference between the short- and long-term periods.

All of the regression analyses were undertaken at hospital level with one observation per hospital per quarter per condition. We weighted each observation of hospital *j* in quarter *q* for condition *c* by a weight equal to the product of the share of each condition of all incentivised and non-incentivised conditions, and the share of each hospital's admissions in a given quarter of all admissions for that condition in the time periods before the introduction of AQ and in the short- and long-term. We used standard errors that were robust to heteroscedasticity. We explored the impact of different variance estimators and the weightings on our results in our sensitivity analyses (see section 3.2.3.5).

The between-region DiD estimation compared the North West with the rest of England. The outcome, y_{jt}, for Trust *j* in quarter t is the percentage of patients that died within 30 days minus the percentage of patients that are expected to die based on the patient-level risk equations. This outcome was modelled as a function of Trust fixed effects (u_j) and time fixed effects (v_t) and a random error term with zero mean (ε_{jt}). We further created variables T_t^1 which equaled 1 if the observation belonged to one of the first 18 months after the start of the AQ programme (short-term, October 2008 – March 2010), T_t^2 which equaled 1 if the observation belonged to one of the next 24 months of the AQ programme (long-term, April 2010 – March 2012), and T_t^3 which equaled 1 if the observation belonged to any period after the AQ programme (short- and long-term combined, October 208 – March 2012).

The interaction of G_j and T_t^1 identifies North West Trusts in the first 18 months after the introduction of AQ. The interaction of G_j and T_t^2 identifies North West Trusts in the next 24 months of AQ. The interaction of G_j and T_t^3 identifies North West Trusts after the introduction of AQ.

To estimate the difference in mortality between the North West and the rest of England before and after AQ we estimated the following equation on the sample of patients with each condition:

$$y_{jt} = \alpha_1 + u_j + \nu_t + \delta_1 G_j \times T_t^1 + \delta_2 G_j \times T_t^2 + \varepsilon_{jt} \qquad (3)$$

In Equation 3, the estimate on the δ_1 coefficient shows how the mean risk-adjusted outcome differs between the AQ and the non-AQ Trusts from before AQ to October 2008 - Mar 2010 (the short-term), conditional on the Trust effects and the time effects. The estimate on δ_2 is the difference in outcome between the AQ and non-AQ Trusts from before AQ to April 2010 - March 2012 (the long-term).

To get an estimate of the difference from the short-term to the long-term, we estimate the following equation:

$$y_{jt} = \alpha_2 + u_j + \nu_t + \delta_2 G_j \times T_t^2 + \delta_3 G_j \times T_t^3 + \varepsilon_{jt} \qquad (4)$$

where the coefficient estimate δ_2 is the difference in the change in mortality between the short-term and the long-term. By definition, the coefficient estimate δ_3 from Equation 4 is equal to δ_1 in Equation 3.

The main effects of G_j are omitted as they are perfectly collinear with the Trust fixed effects (u_j) and the main effects of T_t^x are omitted as they are perfectly collinear with the time fixed effects (v_t).

The time fixed effects are quarters and are captured by 20 dummy variables, with the first quarter (April - June 2007) being the reference category. The coefficients on these dummy variables reveal how the mean value of the outcome changes over time, conditional on the Trust effect.

To estimate the triple-difference model we defined a third dummy variable C_j which takes a value of 1 for the incentivised conditions and a value of 0 for non-incentivised conditions and estimated the following equation:

$$y_{jt} = \alpha_3 + u_j + v_t + \delta_4 G_j \times T_t^1 + \delta_5 G_j \times T_t^2 + \delta_6 C_j \times T_t^1 + \delta_7 C_j \times T_t^2 + \delta_8 G_j \times T_t^1 \times C_j + \delta_9 G_j \times T_t^2 \times C_j + \varepsilon_{jt} \quad (5)$$

where δ_8 and δ_9 are the coefficients of interest capturing the effect of the AQ program on in-hospital mortality for the incentivised conditions in the North West having netted out the effects of changes over time in mortality for the incentivised conditions in the rest of England, changes over time in overall mortality in the North West, and differences in mortality between the incentivised and non-incentivised conditions between the North West and rest of England.

To get an estimate of the difference from the short-term to the long-term in the triple-difference model, we estimated the following equation:

$$y_{jt} = \alpha_4 + u_j + v_t + \delta_{10} G_j \times T_t^2 + \delta_{11} G_j \times T_t^3 + \delta_{12} C_j \times T_t^2 + \delta_{13} C_j \times T_t^3 + \delta_{14} G_j \times T_t^2 \times C_j + \delta_{15} G_j \times T_t^3 \times C_j + \varepsilon_{jt} \quad (6)$$

where the coefficient estimate δ_{14} is the difference in the change in mortality between the short-term and the long-term.

3.2.3.3 Pre-trends tests

A critical assumption of DiD is that changes in the control hospitals act as an appropriate counterfactual for the changes that would have occurred in the treated hospitals in the absence of the programme. We undertook pre-trends tests to confirm that hospitals in the North West did not have a different trend to those in the rest of England prior to the introduction of AQ.

Using data from before the introduction of AQ, we estimated the following regression model for each condition:

$$y_{jt} = \alpha_5 + u_j + \beta \times t + \rho \times G_j \times t + \varepsilon_{jt} \quad (7)$$

in which *t* represents the quarter since the start of the data period, β is an estimate of the quarterly trend in the rest of England and ρ is the difference between the quarterly trend in the North West of England and the quarterly trend in the rest of England.

3.2.3.4 Spillover effects

3.2.3.4.1 Analysis of mortality in two regions that adopted Advancing Quality process measures in the longer term (months 19 to 42)

In the longer term period of the study (months 19 to 42), two other regions outside the North West (the South Central region and the South East Coast regions) adopted the AQ process measures, which had already been used in the North West region, though without the full supporting mechanisms in the AQ programme. In these two regions, (who named their programmes Enhancing Quality and Improving Quality) in the longer-term, there was a financial incentive for performance on the AQ process measures, with money being withheld from hospitals in these regions if they failed to perform to negotiated standards on the AQ process measures.

To test whether the adoption of the AQ measures had an effect on mortality in these two additional regions, we conducted a between region DiD analysis similar to our main analysis, but identifying the South Central and South East Coast regions as a separate group (labelled ‘late-adopter regions’).

3.2.3.4.2 Analysis of improvements in mortality in non-incentivised conditions in hospitals in the North West

To investigate the possibility that the apparent loss of effect of AQ on the incentivised conditions in the North West of England in the long-term was due to improvements in care in the non-incentivised conditions in the intervention hospitals, we first examined the between region DiD separately for each of the non-incentivised conditions. This is similar to the separate analysis of the incentivised conditions reported in the main analysis, but this time carried out for non-incentivised conditions.

We then examined the extent to which patients with the non-incentivised conditions were treated in the same specialties as patients with the incentivised conditions in order to investigate the possibility of spillover quality improvement activities into the non-incentivised conditions in the intervention hospitals. We hypothesised that if positive spillover was to occur, it would most likely affect conditions treated in the same specialties as those exposed to the quality improvement measures in the AQ programme. We therefore examined the extent to which the non-incentivised conditions were treated in the same specialties and by the same specialists as the incentivised conditions. We did this by analysing two fields in the Hospital Episode Statistics data that uniquely identify the specialist team responsible for each patient’s care and under which specialty the lead specialist was employed.

3.2.3.5 Sensitivity analyses

To confirm the robustness of the results of our main analysis, we undertook a number of additional sensitivity analyses:

i. We verified that the conclusions of our main analysis were not sensitive to our use of observations weights or choice of variance estimator.
ii. We repeated the main analysis but using total (in-hospital and out-of-hospital) mortality rates rather than in-hospital mortality rates only to confirm that the results remained stable to using total mortality rates.
iii. We verified that our results were not generated by regression towards the mean by including baseline mortality instead of hospital fixed effects.
iv. We verified that our results were unaffected by the removal of the small group of hospitals that introduced financial incentives for the incentivised or non-incentivised conditions in the longer-term.
v. We confirmed that our results remained stable when analysing 90-day in-hospital mortality rates rather than 30-day in-hospital mortality rates.

3.3 Results

3.3.1 Hospital performance on the incentivised quality measures

The average performance reported by the participating hospitals on all of the quality measures improved in the first 18 months and improved further in the following 24 months, particularly for heart failure and pneumonia (Table 7). Analysis of performance by quarter (Table 8 and Figure 1) showed that rates of improvement slowed over time and, for some quality measures, especially for AMI, plateaued at high levels of achievement towards the end of the period.

Table 7: Average hospital achievement on the incentivised indicators in the North West of England in the first quarter, the last quarter of the short-term period and the last quarter of the long-term period

Condition	Quality indicator	First quarter	Last quarter, short term	Last quarter, long term	Change in short term	Change between short and long term
		Percent of patients			Percentage points	
AMI	Adult smoking cessation advice/counselling	81.1	89.5	95.7	8.4	6.2
	ACEI or ARB for left ventricular systolic dysfunction	96.5	99.2	97.7	2.6	-1.5
	Aspirin at arrival	95.8	98.1	98.7	2.3	0.5
	Aspirin prescribed at discharge	97.8	99.2	99.4	1.3	0.2
	Beta blocker prescribed at discharge	90.6	96.2	98.1	5.7	1.8
	Fibrinolytic therapy received within 30 minutes of hospital arrival	81.1	83.8	92.9	2.6	9.1
Heart failure	Adult smoking cessation advice/counselling	41.8	63.5	78.2	21.7	14.7
	ACEI or ARB for left ventricular systolic dysfunction	89.6	90.6	95.4	1.0	4.8
	Discharge instructions	22.4	37.5	60.7	15.1	23.3
	Left ventricular systolic function assessment	86.2	94.3	91.9	8.1	-2.4
Pneumonia	Adult smoking cessation advice/counselling	36.3	56.4	70.0	20.1	13.6
	Blood cultures performed in ER prior to initial antibiotics received	58.8	74.0	82.4	15.2	8.4
	Initial antibiotic received within six hours of hospital arrival	67.5	71.4	77.4	3.9	5.9
	Initial antibiotic selection in immunocompetent patients	82.1	86.7	92.0	4.6	5.3
	Oxygenation assessment	96.1	98.9	99.6	2.8	0.7

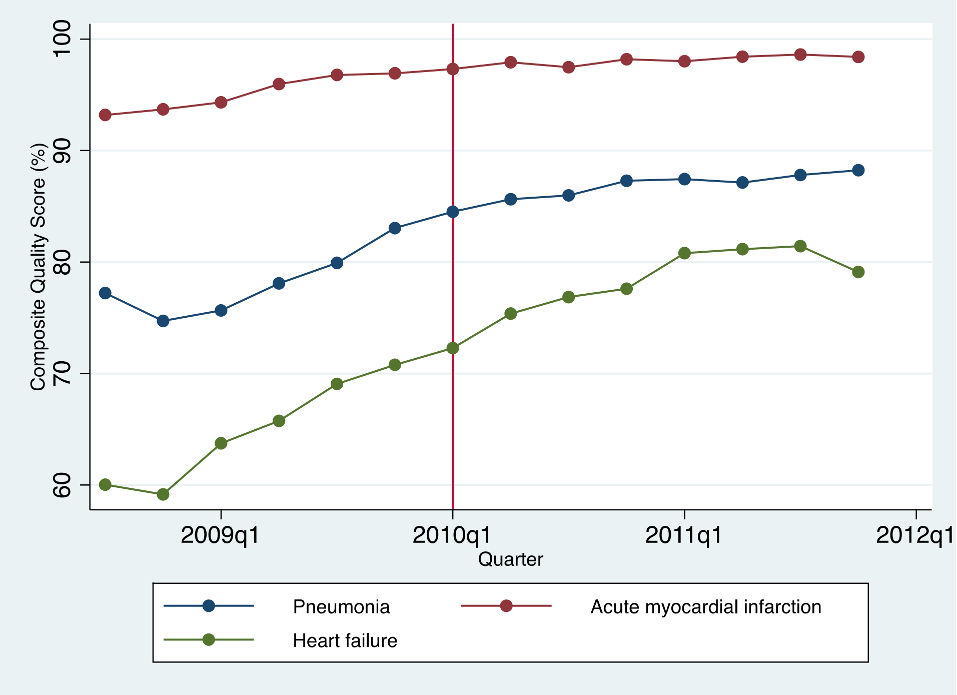
Notes: The table shows the percentage of patients for whom an indicator was met on the incentivised indicators in the North West of England. The short-term period covers the first 18 months of the programme. The long-term period includes months 19-42 of the programme. ER is Emergency Room. ACEI is Angiotensin Converting Enzyme Inhibitor. ARB is Angiotensin Receptor Blocker.

Table 8: Quarterly achievements on the incentivised indicators by hospitals in the North West

Indicator name	Q4 2008	Q1 2009	Q2 2009	Q3 2009	Q4 2009	Q1 2010	Q2 2010	Q3 2010	Q4 2010	Q1 2011	Q2 2011	Q3 2011	Q4 2011	Q1 2012
AMI														
Adult smoking cessation advice/counselling	81.11	82.13	86.3	89.02	90.29	89.46	91.37	92.63	92.6	95.32	92.79	96.16	95.96	95.67
ACEI or ARB for LVSD	96.53	97.27	96.49	97.69	97.27	99.17	97.7	99.33	97.12	96.85	98.66	99.27	97.8	97.7
Aspirin at arrival	95.82	95.41	96.6	97.17	98.19	98.12	98.8	98.6	98.79	98.66	98.93	98.86	99.37	98.67
Aspirin prescribed at discharge	97.82	97.58	97.41	98.2	98.84	99.17	99.36	99.71	99.04	99.46	99.24	99.28	99.25	99.41
Beta Blocker prescribed at discharge	90.56	91.51	92.74	94.55	95.63	96.24	96.39	97.45	96.95	98.08	97.52	97.77	98.19	98.07
Fibrinolytic therapy received within 30 minutes of hospital arrival	81.13	85.25	80.98	89.74	87.17	83.77	83.91	87.39	81.2	88.68	86.25	79.17	84.62	92.86
Heart failure														
Adult smoking cessation advice/counselling	41.79	41.21	60.59	64.41	63.98	63.54	57.07	74.75	74.21	73.23	81.87	85.31	83.65	78.22
ACEI or ARB for LVSD	89.6	86.88	90.56	90.88	91.78	90.6	94.98	93.4	92.41	95.56	94.93	96.04	96.56	95.41
Discharge instructions	22.42	22.82	26.56	31.35	35.55	37.48	42.69	47.01	52.35	52.97	59.29	62.7	62.79	60.73
Evaluation of LVS Function	86.24	85.65	89.49	88.96	91.56	94.32	92.07	94.29	93.38	94.51	95.47	93.72	93.87	91.88
Pneumonia														
Adult smoking cessation advice/counselling	36.32	33.21	40.33	45.58	54.97	56.43	54.25	60.27	59.12	62.08	61.2	67.26	66.27	70.03
Blood cultures performed in ER prior to initial abx received in hospital	58.83	56.49	57.47	62.4	66.49	74.01	79.91	79.25	81.82	81.46	83.39	81.76	83.68	82.42
Initial antibiotic received within 6 hours of hospital arrival	67.51	63.76	63.02	63.33	66.42	71.43	75.37	75.45	73.81	77.28	77.54	76.02	76.5	77.37
Initial antibiotic selection for CAP in immunocompetent patients	82.08	77.96	80.81	83.7	84.14	86.71	86.42	89.88	92.67	92.72	90.86	90.71	92.39	92.04
Oxygenation assessment	96.09	96.07	97.41	98.76	98.52	98.86	99.31	99.2	99.41	99.65	99.72	99.41	99.55	99.56

Notes: The table shows the percentage of patients for whom an indicator was met on the incentivised indicators in the North West of England. ACEI is Angiotensin Converting Enzyme Inhibitor. ARB is Angiotensin Receptor Blocker. LVSD is left ventricular systolic dysfunction.

Figure 1: Average quarterly hospital performance on the incentivised quality measures for each condition



Notes: Vertical line indicates the start of the long-term period (month 19 of the programme). The composite quality score is shown for each quarter, where each indicator is given equal weight in the composite quality score.

3.3.2 Patient characteristics

The characteristics of the patient populations in the North West region and the rest of England were similar before the initiative was introduced, with a slight tendency for patients in the North West to be younger and to have more coexisting conditions (Table 9). Similar changes over the short- and long-term periods in admission volumes and patient characteristics were observed in both regions.

Table 9: Characteristics of patients before and after the introduction of Advancing Quality, in the North West and the rest of England

	North West Region			Rest of England		
	Before AQ	Short term	Long term	Before AQ	Short term	Long term
AMI						
Admissions (n)	19,992	18,804	23,282	104,460	101,765	122,349
Male patients (%)	61.7	61.9	60.3	63.3	63.2	62.7
Patients aged >=75 yr (%)	43.1	43.3	44.7	44	44.9	46.1
Coexisting conditions (average no.)	1.6	1.7	2	1.5	1.7	1.9
Unadjusted mortality in 30 days (%)	11.6	10.5	9.9	10.4	10.1	9.5
Heart Failure						
Admissions (n)	15,295	15,493	20,127	82,847	86,786	118,373
Male patients (%)	53	53.1	52.8	51.3	51.6	52.1
Patients aged >=75 yr (%)	61.4	64	65.7	67.1	68.8	68.8
Coexisting conditions (average no.)	2.3	2.4	2.7	2.2	2.4	2.7
Unadjusted mortality in 30 days (%)	16.4	15.3	14.2	15.3	14.8	13.7
Pneumonia						
Admissions (n)	28,159	36,656	53,180	149,579	196,381	297,999
Male patients (%)	50.4	49.8	50.3	52.2	51	51.2
Patients aged >=75 yr (%)	53.9	55.6	55.1	56.4	58	58.1
Coexisting conditions (average no.)	1.8	2	2.2	1.4	1.6	1.8
Unadjusted mortality in 30 days (%)	26.6	24.7	22.9	25.9	25.1	21.9
Non-incentivised conditions						
Admissions (n)	13,449	14,837	21,975	76,649	84,578	122,503
Male patients (%)	57.4	57.1	55.9	57.2	57	56.7
Patients aged >=75 yr (%)	30.5	33	33	35.1	37.1	37.9
Coexisting conditions (average no.)	1.6	1.7	1.9	1.4	1.6	1.8
Unadjusted mortality in 30 days (%)	13.9	14.1	11.8	12.2	11.8	10.9

Notes: AQ is Advancing Quality. AMI is acute myocardial infarction.

3.3.3 Pre-trends tests

We verified that the trends in mortality were similar in the two regions before the introduction of the AQ programme. We were able to accept the null hypothesis of equal pre-trends in risk-adjusted mortality for each condition. The estimated values for ρ were as follows: AMI -0.34, 95% CI -0.98 to 0.29; heart failure 0.19, 95% CI -0.52 to 0.90; pneumonia -0.23, 95% CI -0.81 to 0.36, non-incentivised conditions -0.66, 95% CI -1.40 to 0.09.

3.3.4 Mortality

Risk-adjusted mortality rates decreased over time in both the North West region and the rest of England, for all of the incentivised and non-incentivised conditions (Table 10). In the short-term, the difference in mortality between the North West region and the rest of England was significantly reduced (Figure 2). In the long-term, mortality rates in the North West remained lower than prior to the introduction of the programme, but the difference between the two regions returned to its pre-intervention level.

The between-region DiD analysis confirmed that the initiative had a significant overall effect on mortality in the short-term (-0.9 percentage points; 95% CI -1.3 to -0.4), comprising a statistically significant reduction in mortality for patients with pneumonia (-1.5 percentage points; 95% CI -2.3 to -0.7) and non-significant reductions for patients with AMI (-0.1 percentage points; 95% CI -0.9 to 0.6) and those with heart failure (-0.2 percentage points; 95% CI -1.1 to 0.7). The triple-difference analysis showed an overall short-term effect of -1.5 percentage points (95% CI -2.6 to -0.5), comprising a statistically significant reduction for patients with pneumonia (-2.2 percentage points; 95% CI -3.4 to -1.0) and non-significant reductions for the other two incentivised conditions.

Between the short-term and long-term periods, risk-adjusted mortality for the incentivised conditions fell by 1.6 percentage points in the North West of England and by 2.3 percentage points in the rest of England. The greater mortality reductions in the rest of England (by 0.7 percentage points; 95% CI 0.3 to 1.2) primarily reflected the reduction in mortality among patients with pneumonia (1.1 percentage points; 95% CI 0.4 to 1.8). However, the reductions in mortality for non-incentivised conditions were larger in the North West region than in the rest of England. The triple-difference analysis showed a larger reduction in mortality (by 1.9 percentage points; 95% CI 1.0 to 2.8) for the rest of England than in the North West between the short-term and long-term periods.

Thus, the short-term improvements in mortality in the North West over those in the rest of England were not maintained. The change in mortality from before the initiative to the end of the long-term period was not statistically significant from zero in either the between-region DiD analysis (-0.1 percentage points; 95% CI -0.6 to 0.3) or the triple-difference analysis (0.4 percentage points; 95% CI -0.6 to 1.3).

Data on out-of-hospital deaths were incomplete for the final three months of the study period and were not included in our main analysis. In sensitivity analyses, we confirmed that our findings were unaffected when we also included out-of-hospital mortality (which was less than 1 percentage point higher than in-hospital mortality for all conditions) (Appendix 3.5.2, Table 16), baseline mortality (Appendix 3.5.3, Table 17), and excluded a small group of control hospitals for whom incentives for the same conditions were introduced in the long-term (Appendix 3.5.4, Table 18). Sensitivity analyses also confirmed our findings when we examined 90-day in-hospital mortality rather than 30-day in-hospital mortality (Appendix 3.5.5, Table 19).

Table 10: Risk-adjusted mortality for the conditions included in Advancing Quality and the non-incentivised conditions examined, before and after the introduction of Advancing Quality

	North West region		**Rest of England**		**Between-region DiD**		**Triple-difference**	
	Rate	Change	Rate	Change	Est.	95% CI	Est.	95% CI
Non-incentivised conditions combined								
Mortality before introduction	14.0		12.3					
Change from before to short-term		*-0.5*		*-1.2*	0.7	(-0.2,1.6)	-	
Mortality in short-term	13.5		11.2					
Change from short-term to long-term		*-2.9*		*-1.7*	-1.2	(-2.0,-0.4)		
Mortality in long-term	10.5		9.5					
Change from before to long-term		*-3.5*		*-2.8*	-0.5	(-1.4,0.3)	-	
Incentivised conditions combined								
Mortality before introduction	20.5		18.9					
Change from before to short-term		*-1.6*		*-0.8*	-0.9	(-1.3,-0.4)	-1.5	(-2.6,-0.5)
Mortality in short-term	18.9		18.1					
Change from short-term to long-term		*-1.8*		*-2.3*	0.7	(0.3,1.2)	1.9	(1.0,2.8)
Mortality in long-term	17.1		15.8					
Change from before to long-term		*-3.4*		*-3.1*	-0.1	(-0.6,0.3)	0.4	(-0.6,1.3)

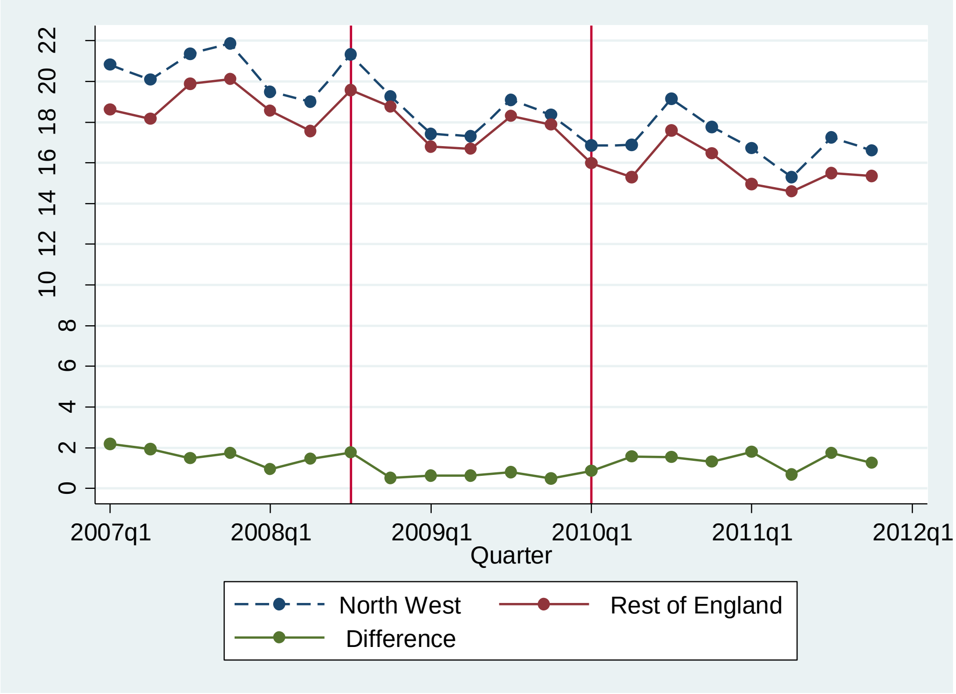
Notes: The short-term period covers the first 18 months of the programme. The long-term period includes months 19-42 of the programme. The between-region DiDs are the changes over time in the North West region minus the changes over time in the rest of England. The triple-difference represents (the change over time in mortality from the conditions included in the programme in the North West region minus the change over time in mortality from the conditions included in the programme in the rest of England) minus (the change over time in mortality from the conditions not included in the programme in the North West region minus the change over time in mortality from the conditions not included in the programme in the rest of England). Estimates are from weighted least-squares regression models that included indicator variables for the quarter of the calendar year during which the admission took place and the admitting hospital; heteroscedasticity-robust standard errors were used in the calculation. The results are robust for other specifications of standard errors and weights (see section Appendix 3.5.1, Table 15). Discrepancies between differences in means and estimated differences are due to rounding and the inclusion of indicator variables for calendar quarter during which the admission took place and admitting hospital in the regression models.

Table 10 continued

Condition, incentivised conditions separately	**North West region**		**Rest of England**		**Between-region DiD**		**Triple-difference**	
	Rate	Change	Rate	Change	Est.	95% CI	Est.	95% CI
AMI								
Mortality before introduction	11.2		10.5					
Change from Before to short-term		*-1.1*		*-0.9*	-0.1	(-0.9,0.6)	-0.8	(-2.0,0.3)
Mortality in short-term	10.0		9.6					
Change from short-term to long-term		*-1.0*		*-1.4*	0.4	(-0.3,1.0)	1.6	(0.5,2.6)
Mortality in long-term	9.0		8.3					
Change from before to long-term		*-2.2*		*-2.3*	0.2	(-0.5,0.9)	0.7	(-0.4,1.8)
Heart failure								
Mortality before introduction	16.9		15.2					
Change from before to short-term		*-1.1*		*-0.9*	-0.2	(-1.1,0.7)	-0.9	(-2.1,0.3)
Mortality in short-term	15.8		14.3					
Change from short-term to long-term		*-1.6*		*-1.6*	0.2	(-0.6,1.0)	1.4	(0.2,2.5)
Mortality in long-term	14.2		12.7					
Change from before to long-term		*-2.7*		*-2.5*	0	(-0.9,0.8)	0.5	(-0.7,1.7)
Pneumonia								
Mortality before introduction	26.9		24.8					
Change from before to short-term		*-2.2*		*-0.7*	-1.5	(-2.3,-0.7)	-2.2	(-3.4,-1.0)
Mortality in short-term	24.7		24.1					
Change from short-term to long-term		*-1.9*		*-3.2*	1.1	(0.4,1.8)	2.3	(1.3,3.4)
Mortality in long-term	22.8		20.9					
Change from before to long-term		*-4.1*		*-3.9*	-0.4	(-1.1,0.3)	0.1	(-1.0,1.2)

Notes: AMI is acute myocardial infarction.

Figure 2: 30-day in-hospital mortality for the conditions included in Advancing Quality



Notes: Vertical lines indicate the start of the short-term period (months 1-18 of the programme) and long-term period (months 19-42 of the programme).

3.3.5 Spillover effects

We considered the possibility that the loss of effect might be due to improvements in care in the control regions, or in non-incentivised conditions in the participating hospitals. These are sometimes termed ‘spillover effects’ (Eijkenaar et al., 2013), (i.e. effects of the intervention that occur outside the targeted clinical or geographical area). We found limited evidence of a positive spillover effect for both possibilities.

In particular, the early results of the AQ programme had been widely disseminated in England, and two regions had adopted a form of the programme’s incentives. Table 11 shows that in the short-term, when AQ was introduced in the North West, but before the process measures were introduced in the late-adopter regions, the reductions in mortality were similar in the rest of England and the late-adopter regions. This was as expected given the lack of incentives in the late-adopter regions in the short-term.

In the longer-term, when the AQ process measures were used in the late-adopter regions, reductions in 30 day in-hospital mortality were greater in the late-adopter regions than in the rest of England, but this was only statistically significant for AMI (-0.7 percentage points; 95% CI -1.3 to -0.1). This suggested that the adoption of the AQ indicators in the two late-adopter regions was associated with a reduction in mortality for AMI.

We also found limited evidence of a positive spillover effect within the AQ hospitals. Within AQ hospitals, non-incentivised conditions which were treated by specialists who also treated patients with the incentivised conditions showed the largest reductions in mortality amongst the non-incentivised conditions in the long-term. The results show that patients with incentivised and some non-incentivised conditions were indeed being treated in the same specialties and by the same specialist teams (Table 12 and Table 13). Specifically, Table 12 shows that, at specialty level, 41% to 51% of patients with the incentivised conditions were treated in General Medicine. This is also the specialty where 48% of patients with acute renal failure and 57% of patients with alcoholic liver disease were treated. These are the two conditions that showed the greater reductions in mortality from the short to the long-term in the North West (Table 14). There was less overlap between the specialties treating the other non-incentivised conditions, which were also those with non-significant reductions in mortality in the later period.

We also found overlap at the lower level of individual specialists’ teams that treated patients with incentivised and non-incentivised conditions. Table 13 shows that 55% of specialists treating at least one AMI patient had treated at least one patient with acute renal failure. It can be seen, that while this level of overlap is not uncommon across all incentivised and non-incentivised conditions, the proportion of specialists with a substantial workload of both patients with incentivised and patients with non-incentivised conditions (defined here as treating at least 20% of patients with either condition (of the total of any two conditions)) remains high for the two non-incentivised conditions with statistically significantly larger mortality reductions in the North West, and is lower for patients with other non-incentivised conditions.

In summary, these findings can be interpreted as modest evidence of potential mechanisms through with AQ could have affected the care of patients for some of the non-incentivised conditions (in particular acute renal failure and alcoholic liver disease).

Table 11: Risk-adjusted mortality for the conditions included in Advancing Quality and the non-incentivised conditions examined, before and after the introduction of Advancing Quality in the North West Region of England and the adaption of the programme's quality metrics in the late-adopter regions

							Between-region Difference-in-Differences			
	North West region		**Late-adopter regions**		**Rest of England**		**North West versus rest of England**		**Late-adopters versus rest of England**	
Non-incentivised conditions	*Rate*	*Change*	*Rate*	*Change*	*Rate*	*Change*	*Est.*	*C.I.*	*Est.*	*C.I.*
Mortality before Introduction	14.0		11.8		12.5					
Change from Before to Short-Term		*-0.5*		*-0.9*		*-1.2*	0.8	(-0.1,1.7)	0.3	(-0.4,1.0)
Mortality after Introduction (Short term)	13.5		10.9		11.3					
Change from Short-Term to Long-Term		*-2.9*		*-1.9*		*-1.6*	-1.3	(-2.0,-0.5)	-0.2	(-0.8,0.4)
Mortality after Introduction (Long term)	10.5		9		9.7					
Change from Before to Long-Term		*-3.5*		*-2.8*		*-2.8*	-0.5	(-1.3,0.4)	0.1	(-0.5,0.8)
Incentivised conditions combined										
Mortality before Introduction	20.5		18.6		19					
Change from Before to Short-Term		*-1.6*		*-0.8*		*-0.8*	-0.9	(-1.4,-0.4)	-0.1	(-0.6,0.3)
Mortality after Introduction (Short term)	18.9		17.8		18.2					
Change from Short-Term to Long-Term		*-1.8*		*-2.6*		*-2.2*	0.7	(0.2,1.1)	-0.2	(-0.6,0.2)
Mortality after Introduction (Long term)	17.1		15.2		16					
Change from Before to Long-Term		*-3.4*		*-3.4*		*-3.0*	-0.2	(-0.7,0.2)	-0.3	(-0.8,0.1)
AMI										
Mortality before Introduction	11.2		10		10.7					
Change from Before to Short-Term		*-1.1*		*-0.6*		*-1.0*	-0.1	(-0.8,0.7)	0.3	(-0.4,0.9)
Mortality after Introduction (Short term)	10.0		9.4		9.7					
Change from Short-Term to Long-Term		*-1.0*		*-1.9*		*-1.1*	0.2	(-0.5,0.8)	-0.7	(-1.3,-0.1)
Mortality after Introduction (Long term)	9.0		7.5		8.6					
Change from Before to Long-Term		*-2.2*		*-2.5*		*-2.1*	0.1	(-0.7,0.8)	-0.5	(-1.1,0.2)
Heart failure										
Mortality before Introduction	16.9		15		15.3					
Change from Before to Short-Term		*-1.1*		*-0.7*		*-1.1*	-0.1	(-1.0,0.8)	0.3	(-0.6,1.1)
Mortality after Introduction (Short term)	15.8		14.3		14.2					
Change from Short-Term to Long-Term		*-1.6*		*-2.1*		*-1.3*	0	(-0.8,0.8)	-0.7	(-1.4,0.1)
Mortality after Introduction (Long term)	14.2		12.2		12.9					
Change from Before to Long-Term		*-2.7*		*-2.8*		*-2.4*	-0.1	(-1.0,0.7)	-0.4	(-1.2,0.4)
Pneumonia										
Mortality before Introduction	26.9		24.6		24.9					
Change from Before to Short-Term		*-2.2*		*-1.1*		*-0.6*	-1.7	(-2.5,-0.9)	-0.5	(-1.2,0.2)
Mortality after Introduction (Short term)	24.7		23.5		24.3					
Change from Short-Term to Long-Term		*-1.9*		*-3.0*		*-3.2*	1.2	(0.5,1.9)	0.2	(-0.4,0.8)
Mortality after Introduction (Long term)	22.8		20.5		21.1					
Change from Before to Long-Term		*-4.1*		*-4.1*		*-3.8*	-0.5	(-1.2,0.3)	-0.2	(-0.9,0.4)

Note: Late-adopters are the two regions in England (South Central and South East Coast) that formally adopted the Advancing Quality metrics in months 19-42. The short-term period covers the first 18 months of the programme. The long-term period includes months 19-42 of the programme. The between-region differences-in-differences are the changes over time in the North West or late-adopter regions minus the changes over time in the rest of England. Estimates are from weighted least-squares regressions that include indicator variables for quarter of admission and admitting hospital using heteroscedasticity robust standard errors. AMI is acute myocardial infarction.

Table 12: Analysis of overlap between care of patients with incentivised and non-incentivised conditions in the same specialties, percentage of patients treated under each specialty

	Incentivised conditions			Non-incentivised conditions				
	AMI	Heart failure	Pneumonia	Acute renal failure	Alcoholic liver disease	Duodenal ulcer	Intracranial injury	Paralytic ileus & intestinal obstruction without hernia
Specialties in which incentivised conditions are concentrated								
General medicine	41	48	53	48	57	38	18	9
Geriatric medicine	9	16	17	15	6	6	6	3
Cardiology	35	16	3	2	2	1	1	0
Respiratory medicine	3	5	8	3	3	2	1	1
Specialties in which non-incentivised conditions are concentrated								
Gastroenterology	3	4	4	4	18	20	2	2
Accident and emergency (A&E)	2	2	3	3	3	2	24	3
General surgery	0	1	1	4	3	24	11	74
Neurosurgery	0	0	0	0	0	0	18	0
Nephrology	1	1	1	8	1	1	0	0
Trauma and orthopaedics	0	0	0	0	0	0	9	0
Total	93	92	90	87	93	93	90	93

Notes: Figures are presented for specialties treating at least 5% of patients for one of the incentivised or non-incentivised conditions. AMI is acute myocardial infarction.

Table 13: Analysis of overlap between care of patients with incentivised and non-incentivised conditions by the same specialists, percentage of specialists treating at least 1 patient with an incentivised condition who treated at least 1 patient with a non-incentivised condition

	Acute renal failure	Alcoholic liver disease	Duodenal ulcer	Intracranial injury	Paralytic ileus & intestinal obstruction without hernia
AMI	55 [38]	48 [25]	43 [17]	42 [16]	39 [12]
Heart failure	55 [32]	47 [19]	42 [13]	40 [10]	38 [8]
Pneumonia	53 [16]	38 [9]	40 [13]	38 [10]	42 [9]

[ ] the percentage of specialists treating at least 20% of patients from either condition (of the total of two conditions)

AMI is acute myocardial infarction.

Table 14: Risk-adjusted mortality for the non-incentivised conditions separately, before and after the introduction of Advancing Quality

	North West region		Rest of England		Between region DiD	
	Rate	Change	Rate	Change	Est.	C.I.
Acute renal failure						
Mortality before introduction	20.1		17.3			
Change from before to short-term		*-1.5*		*-1.6*	0.2	(-1.8,2.2)
Mortality in short-term	18.6		15.7			
Change from short-term to long-term		*-5.4*		*-2.5*	-2.7	(-4.4,-1.1)
Mortality in long-term	13.2		13.2			
Change from before to long-term		*-6.9*		*-4.1*	-2.5	(-4.4,-0.7)
Alcoholic liver disease						
Mortality before introduction	14.4		14.4			
Change from before to short-term		*0.5*		*-0.9*	1.7	(-0.2,3.6)
Mortality in short-term	14.9		13.5			
Change from short-term to long-term		*-3.2*		*-1.3*	-2.0	(-3.7,-0.3)
Mortality in long-term	11.7		12.2			
Change from before to long-term		*-2.7*		*-2.2*	-0.4	(-2.2,1.5)
Duodenal ulcer						
Mortality before introduction	7		7.5			
Change from before to short-term		*-0.6*		*-1.6*	1.1	(-1.0,3.1)
Mortality in short-term	6.4		5.9			
Change from Short-Term to Long-Term		*-1.3*		*-1.2*	0.0	(-1.9,1.9)
Mortality in long-term	5.1		4.7			
Change from before to long-term		*-1.9*		*-2.8*	1.1	(-0.8,2.9)
Intracranial injury						
Mortality before introduction	14.2		12.7			
Change from before to short-term		*-1.0*		*-1.4*	0.7	(-1.5,2.9)
Mortality in short-term	13.2		11.3			
Change from short-term to long-term		*-1.9*		*-1.4*	-0.2	(-2.1,1.8)
Mortality in long-term	11.3		9.9			
Change from before to long-term		*-2.9*		*-2.8*	0.6	(-1.5,2.6)
Paralytic ileus & intestinal obstruction without hernia						
Mortality before introduction	9.4		7.8			
Change from before to short-term		*-0.7*		*-0.6*	-0.2	(-1.6,1.3)
Mortality in short-term	8.7		7.2			
Change from short-term to long-term		*-0.9*		*-1.2*	0.5	(-0.8,1.7)
Mortality in long-term	7.8		6			
Change from before to long-term		*-1.6*		*-1.8*	0.3	(-1.0,1.6)

Notes: The short-term covers the first 18 months of the programme. The long-term includes month 19-42. The between region DiDs are the changes over time in the North West region minus the changes over time in the rest of England.

3.4 Discussion

Earlier work showed that the introduction of a P4P programme for all hospitals in the North West region of England was associated with a significant reduction in mortality for the conditions linked to incentives in the first 18 months of the programme. The current analysis shows that in the following 24 months, although mortality in this region continued to decline, the decline for the conditions linked to incentives were smaller. In addition, as compared with mortality in the period before the initiative was introduced, there was no longer a significant difference in the decline in mortality between the North West region and the rest of England. These findings were due in part to the significant reductions in mortality between the short-term and long-term periods for the incentivised conditions in the rest of England, which were not matched in the North West. For the non-incentivised conditions, we found no significant difference between regions in the changes in mortality in the short-term, but reductions in mortality between the short-term and long-term periods were significantly larger in the North West region than in the rest of England.

We considered several explanations for these effects. The first is that the incentives were no longer effective, possibly because of the change in incentive structure from bonuses to penalties. The continued improvement in performance on the incentivised quality measures in the AQ hospitals suggests that the incentives may still have been effective, but this information is not collected for control hospitals. It is also possible that initial reductions in mortality reflected the effect of the intervention on the most severely ill patients, leaving less room for subsequent reductions in mortality.

Another possible explanation is that there were positive spillovers in the quality of care from participating to non-participating hospitals and, within participating hospitals, from incentivised to non-incentivised conditions. We found some modest evidence for both types of spillover effects. After the early results showed reductions in mortality, two other regions in English introduced financial incentives for the AQ measures during the long-term period of our study, albeit with none of the supporting mechanisms of the AQ programme. As compared with the regions that did not introduce the incentives, these two regions had a larger long-term reduction in mortality for the conditions linked to incentives, although the reduction was significant only for AMI. Our finding that among conditions not linked to incentives, the largest reductions in mortality were for conditions treated by the same specialists who were treating conditions that were linked to incentives also lends some support to the hypothesis that there may have been positive spillover effects in the AQ hospitals. Spillover effects have previously been considered in the literature (Mullen et al., 2010; Sutton et al., 2010). They may be regarded as positive consequences of a successful quality improvement programme, or as negative consequences if their effect is to reduce the quality of care for non-incentivised conditions (Doran et al., 2011; Eijkenaar et al., 2013). Further exploration of positive and negative spillover effects from P4P initiatives will be important.

In conclusion, although short-term improvements in the quality measures for conditions related to incentives were sustained in the long-term, our analyses provide no evidence that the incentives have a long-term effect on 30-day mortality. Possible explanations for this result are that the effects of P4P were temporal, early effects on outcomes were easier to achieve (i.e. they represented low-hanging fruit), the nature of the incentive was changed (from bonuses for good performance to the withholding of a percentage of reimbursement for poor performance), and there were unintended but desirable spillover effects into other geographical and clinical areas.

3.5 Appendix

3.5.1 The effects of weighting and clustering on standard errors

Table 15: Risk-adjusted mortality for the conditions included in Advancing Quality and the non-incentivised conditions examined, before and after the introduction of Advancing Quality, with different weight and standard error specifications

	Unweighted regressions				Weighted regressions[a]			
	Coef	95% Confidence Interval			Coef	95% Confidence Interval		
		Std. errors				Std. errors		
		OLS	Robust[b]	Cluster[c]		OLS	Robust[b]	Cluster[c]
Non-incentivised conditions								
Change from Before to after P4P (Short-Term)	**0.38**	(-0.63,1.38)	(-0.66,1.41)	(-0.79,1.55)	**0.70**	(-0.06,1.46)	(-0.19,1.58)	(-0.28,1.68)
Change from Short-Term to Long-Term	**-1.25**	(-2.19,-0.30)	(-2.14,-0.35)	(-2.31,-0.18)	**-1.21**	(-1.97,-0.46)	(-1.98,-0.44)	(-2.29,-0.13)
Change from Before to after P4P (Long-Term)	**-0.87**	(-1.81,0.07)	(-1.84,0.11)	(-2.08,0.34)	**-0.51**	(-1.27,0.24)	(-1.36,0.33)	(-1.72,0.69)
Incentivised conditions (total)								
Change from Before to after P4P (Short-Term)	**-0.79**	(-1.36,-0.22)	(-1.33,-0.25)	(-1.38,-0.19)	**-0.86**	(-1.29,-0.42)	(-1.34,-0.37)	(-1.45,-0.27)
Change from Short-Term to Long-Term	**0.71**	(0.18,1.25)	(0.24,1.18)	(0.05,1.38)	**0.71**	(0.27,1.15)	(0.27,1.16)	(0.04,1.38)
Change from Before to after P4P (Long-Term)	**-0.08**	(-0.61,0.46)	(-0.60,0.45)	(-0.85,0.70)	**-0.15**	(-0.58,0.29)	(-0.61,0.31)	(-0.96,0.67)
AMI								
Change from Before to after P4P (Short-Term)	**-0.36**	(-1.24,0.53)	(-1.27,0.55)	(-1.38,0.67)	**-0.13**	(-0.82,0.56)	(-0.85,0.59)	(-1.11,0.86)
Change from Short-Term to Long-Term	**0.67**	(-0.16,1.50)	(-0.05,1.39)	(-0.37,1.71)	**0.35**	(-0.34,1.05)	(-0.30,1.00)	(-0.58,1.28)
Change from Before to after P4P (Long-Term)	**0.31**	(-0.52,1.14)	(-0.61,1.23)	(-0.95,1.58)	**0.22**	(-0.47,0.92)	(-0.50,0.95)	(-0.94,1.38)
Heart failure								
Change from Before to after P4P (Short-Term)	**-0.18**	(-1.24,0.89)	(-1.13,0.78)	(-1.12,0.77)	**-0.20**	(-1.06,0.66)	(-1.06,0.66)	(-1.03,0.63)
Change from Short-Term to Long-Term	**0.34**	(-0.66,1.34)	(-0.55,1.23)	(-0.73,1.41)	**0.17**	(-0.71,1.04)	(-0.65,0.99)	(-0.77,1.10)
Change from Before to after P4P (Long-Term)	**0.17**	(-0.84,1.17)	(-0.77,1.10)	(-1.03,1.36)	**-0.03**	(-0.90,0.84)	(-0.86,0.80)	(-1.14,1.07)
Pneumonia								
Change from Before to after P4P (Short-Term)	**-1.83**	(-2.83,-0.83)	(-2.73,-0.93)	(-2.86,-0.80)	**-1.52**	(-2.24,-0.81)	(-2.30,-0.75)	(-2.46,-0.59)
Change from Short-Term to Long-Term	**1.12**	(0.19,2.05)	(0.31,1.93)	(-0.18,2.42)	**1.14**	(0.41,1.86)	(0.43,1.84)	(0.02,2.25)
Change from Before to after P4P (Long-Term)	**-0.71**	(-1.64,0.22)	(-1.55,0.13)	(-2.08,0.65)	**-0.39**	(-1.11,0.33)	(-1.10,0.33)	(-1.70,0.92)

Notes: a) In the weighted regressions, the weight of each hospital-quarter-condition observation is the product of the share of each condition of all incentivised and control non-incentivised conditions, and the share of each hospitals' admissions in a given quarter of all admissions for that condition in the time periods before the introduction of Advancing Quality and in the short- and long-term.

b) The standard errors robust specifications use the Huber/White/sandwich variance estimator robust to unspecified heteroscedasticity.

c)The cluster specifications use the clustered sandwich estimator robust to unspecified heteroscedasticity and within-hospital variation not captures by the hospital fixed effects.

AMI is acute myocardial infarction

3.5.2 Analysis using total (both in- and out-of-hospital) mortality rates

Table 16: Risk-adjusted total mortality for the conditions included in Advancing Quality and the non-incentivised conditions examined, before and after the introduction of Advancing Quality

	North West region		**Rest of England**		**Between region DiD**	
	Rate	Change	Rate	Change	Est.	95% CI
Non-incentivised conditions combined						
Mortality before introduction	14.9		13.1			
Change from before to short-term		*-0.6*		*-1.1*	0.6	(-0.3,1.6)
Mortality in short-term	14.3		12			
Change from short-term to long-term		*-4.9*		*-3.7*	-1.3	(-2.3,-0.2)
Mortality in long-term	9.4		8.3			
Change from before to long-term		*-5.5*		*-4.8*	-0.6	(-1.8,0.6)
Incentivised conditions combined						
Mortality before introduction	21.1		19.4			
Change from before to short-term		*-1.5*		*-0.8*	-0.7	(-1.3,-0.1)
Mortality in short-term	19.6		18.6			
Change from short-term to long-term		*-4.6*		*-5.3*	0.6	(-0.2,1.4)
Mortality in long-term	15		13.3			
Change from before to long-term		*-6.1*		*-6.1*	-0.1	(-1.1,0.9)
AMI						
Mortality before introduction	11.5		10.9			
Change from before to short-term		*-1.2*		*-1.0*	-0.1	(-1.1,0.9)
Mortality in short-term	10.3		9.9			
Change from short-term to long-term		*-2.2*		*-2.8*	0.4	(-0.5,1.3)
Mortality in long-term	8.1		7.1			
Change from before to long-term		*-3.4*		*-3.8*	0.3	(-0.8,1.4)
Heart failure						
Mortality before introduction	17.7		15.9			
Change from before to short-term		*-1.0*		*-0.9*	-0.1	(-0.9,0.8)
Mortality in short-term	16.7		15			
Change from short-term to long-term		*-4.0*		*-3.8*	0	(-1.0,1.0)
Mortality in long-term	12.7		11.2			
Change from before to long-term		*-5.0*		*-4.7*	-0.1	(-1.4,1.1)
Pneumonia						
Mortality before introduction	27.5		25.2			
Change from before to short-term		*-1.9*		*-0.5*	-1.3	(-2.3,-0.4)
Mortality in short-term	25.6		24.7			
Change from short-term to long-term		*-5.9*		*-7.2*	1	(-0.1,2.1)
Mortality in long-term	19.7		17.5			
Change from before to long-term		*-7.8*		*-7.7*	-0.4	(-1.9,1.1)

Notes: The total mortality rate includes both in and out of hospital deaths, but the out of hospital mortality data is incomplete for the 3 final months of the study which explains the large decrease in mortality in the long-term.

The short-term period covers the first 18 months of the programme. The long-term period includes months 19-42 of the programme.

The between-region DiDs are the changes over time in the North West region minus the changes over time in the rest of England.

Estimates are from weighted least-squares regression models that include indicator variables for quarter of admission and admitting hospital using heteroscedasticity robust standard errors.

AMI is acute myocardial infarction.

3.5.3 Effects of including baseline mortality

Table 17: Risk-adjusted mortality for the conditions included in Advancing Quality and the non-incentivised conditions examined, before and after the introduction of Advancing Quality, including baseline mortality

	Between-Region Difference in Differences		Between-Region Difference in Differences with baseline		Triple Differences		Triple Differences with baseline	
	Est.	*C.I.*			*Est.*	*C.I.*	*Est.*	*C.I.*
Non-Incentivised conditions								
Change from Before to Short-Term	0.7	(-0.2,1.6)	1.1	(0.5,1.8)				
Change from Short-Term to Long-Term	-1.2	(-2.0,-0.4)	-1.3	(-2.1,-0.5)				
Change from Before to Long-Term	-0.5	(-1.4,0.3)	-0.2	(-0.7,0.4)				
Incentivised conditions combined								
Change from Before to Short-Term	-0.9	(-1.3,-0.4)	-0.4	(-0.7,-0.0)	-1.5	(-2.6,-0.5)	-1.3	(-2.0,-0.5)
Change from Short-Term to Long-Term	0.7	(0.3,1.2)	0.8	(0.3,1.2)	1.9	(1.0,2.8)	2.1	(1.1,3.0)
Change from Before to Long-Term	-0.1	(-0.6,0.3)	0.4	(0.1,0.7)	0.4	(-0.6,1.3)	0.8	(0.2,1.4)
Acute myocardial infarction								
Change from Before to Short-Term	-0.1	(-0.9,0.6)	0	(-0.5,0.6)	-0.8	(-2.0,0.3)	-1	(-1.8,-0.2)
Change from Short-Term to Long-Term	0.4	(-0.3,1.0)	0.4	(-0.3,1.1)	1.6	(0.5,2.6)	1.7	(0.6,2.7)
Change from Before to Long-Term	0.2	(-0.5,0.9)	0.4	(-0.1,0.9)	0.7	(-0.4,1.8)	0.6	(-0.1,1.4)
Heart failure								
Change from Before to Short-Term	-0.2	(-1.1,0.7)	0.4	(-0.2,1.1)	-0.9	(-2.1,0.3)	-0.5	(-1.4,0.4)
Change from Short-Term to Long-Term	0.2	(-0.6,1.0)	0.2	(-0.7,1.1)	1.4	(0.2,2.5)	1.5	(0.3,2.7)
Change from Before to Long-Term	0	(-0.9,0.8)	0.6	(0.0,1.2)	0.5	(-0.7,1.7)	1	(0.2,1.8)
Pneumonia								
Change from Before to Short-Term	-1.5	(-2.3,-0.7)	-0.9	(-1.5,-0.4)	-2.2	(-3.4,-1.0)	-1.7	(-2.6,-0.9)
Change from Short-Term to Long-Term	1.1	(0.4,1.8)	1.2	(0.4,2.0)	2.3	(1.3,3.4)	2.5	(1.4,3.6)
Change from Before to Long-Term	-0.4	(-1.1,0.3)	0.3	(-0.3,0.8)	0.1	(-1.0,1.2)	0.8	(0.0,1.5)

Notes: The short-term period covers the first 18 months of the programme. The long-term includes months 19-42 of the programme. The between-region DiDs are the changes over time in the North West region minus the changes over time in the rest of England. Estimates are from weighted least-squares regressions that contain indicator variables for quarter of admission and admitting hospital using heteroscedasticity robust standard errors. AMI is acute myocardial infarction

3.5.4 Analysis excluding hospitals with any financial incentives for the conditions incentivised by Advancing Quality or the non-incentivised conditions examined

We repeated the main analysis without the small group of hospitals that introduced financial incentives under the Commissioning for Quality and Innovation (CQUIN) framework for the incentivised or non-incentivised conditions during the long-term period. This analysis confirms that the results are unaffected by the removal of this small group. This small group of hospitals did not incentivise any of the conditions under examination during the short-term period, and so were included in the main analysis to allow direct comparisons between the results estimated for the short- and long-term periods, and to the results of the previous short-term evaluations (Chapter 2).

Table 18: Risk-adjusted mortality for the conditions included in Advancing Quality and the non-incentivised conditions examined, before and after the introduction of Advancing Quality, excluding hospitals with incentives for any of these conditions during the period of evaluation

	North West region		Rest of England		Between-region DiD	
	Rate	Change	Rate	Change	Est.	C.I.
Non-incentivised conditions combined						
Mortality before introduction	13.9		11.7			
Change from before to short-term		*-0.3*		*-0.6*	0.7	(-0.2,1.5)
Mortality in short-term	13.6		11.1			
Change from short-term to long-term		*-2.5*		*-1.3*	-1.2	(-1.9,-0.4)
Mortality in long-term	11.1		9.8			
Change from before to long-term		*-2.8*		*-1.9*	-0.5	(-1.4,0.3)
Incentivised conditions combined						
Mortality before introduction	24.5		19.8			
Change from before to short-term		*-1.1*		*0.1*	-0.7	(-1.2,-0.2)
Mortality in short-term	23.4		19.9			
Change from short-term to long-term		*-1.3*		*-1.7*	0.6	(0.2,1.0)
Mortality in long-term	22.1		18.2			
Change from before to long-term		*-2.4*		*-1.6*	-0.1	(-0.5,0.4)
AMI						
Mortality before introduction	14.7		9.4			
Change from before to short-term		*-0.7*		*-0.9*	0.1	(-0.6,0.8)
Mortality in short-term	14		8.5			
Change from short-term to long-term		*-1.0*		*-0.9*	0.1	(-0.5,0.8)
Mortality in long-term	13		7.6			
Change from before to long-term		*-1.7*		*-1.8*	0.2	(-0.5,1.0)
Heart failure						
Mortality before introduction	24.1		16.6			
Change from before to short-term		*-0.4*		*-0.6*	0	(-0.9,0.8)
Mortality in short-term	23.7		16			
Change from short-term to long-term		*-2.7*		*-1.4*	0.1	(-0.7,0.9)
Mortality in long-term	21		14.6			
Change from before to long-term		*-3.1*		*-2.0*	0.1	(-0.7,0.9)
Pneumonia						
Mortality before introduction	26.4		24.1			
Change from before to short-term		*-1.7*		*-0.4*	-1.4	(-2.2,-0.6)
Mortality in short-term	24.7		23.7			
Change from short-term to long-term		*-1.5*		*-2.7*	0.9	(0.3,1.6)
Mortality in long-term	23.2		21			
Change from before to long-term		*-3.2*		*-3.1*	-0.5	(-1.2,0.2)

Notes: The short-term period covers the first 18 months of the programme. The long-term period includes months 19-42 of the programme. The between-region DiDs are changes over time in the North West region minus changes over time in the rest of England. Estimates are from weighted least-squares regressions that include indicator variables for quarter of admission and admitting hospital using heteroscedasticity robust standard errors. AMI is acute myocardial infarction.

3.5.5 Effects of using 90-day mortality

Table 19: Risk-adjusted mortality for the conditions included in Advancing Quality and the non-incentivised conditions examined, before and after the introduction of Advancing Quality, using 90-day in-hospital mortality

	North West Region		Rest of England		Between-Region Difference-in-Differences		Triple Difference	
	Rate	*Change*	*Rate*	*Change*	*Est.*	*C.I.*	*Est.*	*C.I.*
Non-Incentivised conditions								
Mortality before Introduction	17.9		15.8					
Change from Before to Short-Term		*-1.0*		*-1.5*	0.6	(-0.5,1.6)		
Mortality after Introduction (Short term)	16.9		14.3					
Change from Short-Term to Long-Term		*-3.6*		*-2.3*	-1.3	(-2.2,-0.4)		
Mortality after Introduction (Long term)	13.2		12.0					
Change from Before to Long-Term		*-4.7*		*-3.8*	-0.7	(-1.7,0.3)		
Incentivised conditions combined								
Mortality before Introduction	25.1		23.1					
Change from Before to Short-Term		*-2.0*		*-1.1*	-1	(-1.5,-0.4)	-1.5	(-2.6,-0.4)
Mortality after Introduction (Short term)	23.1		21.9					
Change from Short-Term to Long-Term		*-2.8*		*-3.2*	0.7	(0.2,1.2)	2	(0.9,3.0)
Mortality after Introduction (Long term)	20.4		18.7					
Change from Before to Long-Term		*-4.8*		*-4.3*	-0.3	(-0.8,0.2)	0.4	(-0.6,1.5)

Notes: The short-term period covers the first 18 months of the programme. The long-term includes months 19-42 of the programme.
The between-region difference-in-differences are the changes over time in the North West region minus the change over time in the rest of England. The triple-difference represents (the change over time in mortality from the conditions included in the programme in the North West region minus the change over time in mortality from the conditions included in the programme in the rest of England) minus (the change over time in mortality from the conditions not included in the programme in the North West region minus the change over time in mortality from the conditions not included in the programme in the rest of England).
Estimates are from weighted least-squares regressions that include indicator variables for quarter of admission and admitting hospital using heteroscedasticity robust standard errors.

Table 19 continued

	North West Region		Rest of England		Between-Region Difference-in-Differences		Triple Difference	
	Rate	*Change*	*Rate*	*Change*	*Est.*	*C.I.*	*Est.*	*C.I.*
Acute myocardial infarction								
Mortality before Introduction	13.8		12.7					
Change from Before to Short-Term		*-1.3*		*-1.0*	-0.1	(-0.9,0.6)	-0.7	(-1.9,0.6)
Mortality after Introduction (Short term)	12.5		11.7					
Change from Short-Term to Long-Term		*-1.7*		*-1.9*	0.3	(-0.4,1.1)	1.6	(0.4,2.8)
Mortality after Introduction (Long term)	10.8		9.8					
Change from Before to Long-Term		*-3.0*		*-2.9*	0.2	(-0.6,1.0)	0.9	(-0.4,2.2)
Heart failure								
Mortality before Introduction	22.5		20.3					
Change from Before to Short-Term		*-1.6*		*-1.2*	-0.4	(-1.5,0.6)	-1	(-2.4,0.5)
Mortality after Introduction (Short term)	20.9		19.1					
Change from Short-Term to Long-Term		*-2.4*		*-2.6*	0.3	(-0.7,1.3)	1.6	(0.2,2.9)
Mortality after Introduction (Long term)	18.5		16.5					
Change from Before to Long-Term		*-4.0*		*-3.8*	-0.1	(-1.2,0.9)	0.6	(-0.8,2.0)
Pneumonia								
Mortality before Introduction	32.2		29.6					
Change from Before to Short-Term		*-2.7*		*-1.1*	-1.6	(-2.4,-0.8)	-2.2	(-3.5,-0.9)
Mortality after Introduction (Short term)	29.5		28.5					
Change from Short-Term to Long-Term		*-3.1*		*-4.2*	1	(0.3,1.8)	2.3	(1.1,3.5)
Mortality after Introduction (Long term)	26.3		24.3					
Change from Before to Long-Term		*-5.9*		*-5.4*	-0.6	(-1.4,0.2)	0.1	(-1.1,1.4)

Notes: AMI is acute myocardial infarction

4. Using survival analysis to improve estimates of life year gains in policy evaluations

Abstract

Background: Policy evaluations taking a lifetime horizon have converted estimated changes in short-term mortality to expected life year gains using general population life expectancy. However, the life expectancy of the affected patients may differ from the general population. In trials, survival models are commonly used to extrapolate life year gains. The objective was to demonstrate the feasibility and materiality of using parametric survival models to extrapolate future survival in health care policy evaluations.

Methods: We used our previous cost-effectiveness analysis of a pay-for-performance programme as a motivating example. We first used the cohort of patients admitted prior to the introduction of the programme to compare three methods for estimating remaining life expectancy. We then used a difference-in-differences framework to estimate the life year gains associated with the programme using general population life expectancy and survival models. Patient-level data from Hospital Episode Statistics was utilised for patients admitted to hospitals in England for pneumonia between 1st April 2007 and 31st March 2008 and between 1st April 2009 and 31st March 2010, and linked to death records for the period from 1st April 2007 to 31st March 2011.

Results: In our cohort of patients, using parametric survival models rather than general population life expectancy figures reduced the estimated mean life years remaining by 30% (9.19 versus 13.15 years, respectively). However, the estimated mean life year gains associated with the programme are larger using survival models (0.380 years) compared to using general population life expectancy (0.154 years).

Conclusions: Using general population life expectancy to estimate the impact of health care policies can overestimate life expectancy but underestimate the impact of policies on life year gains. Using a longer follow-up period improved the accuracy of estimated survival and programme impact considerably.

This chapter is published in *Medical Decision Making.*

Citation: Meacock R, Sutton M, Kristensen S, Harrison M. (2017). Using survival analysis to improve estimates of life year gains in policy evaluations. *Medical Decision Making,* 37, 415-426.

4.1 Introduction

The effects of health care policies and programmes should be evaluated in terms of their impact on health outcomes, as is now standard practice for all new health care technologies. This impact can be comprised of effects on both the quality and length of life. Length of life is a key outcome for cost-effectiveness analysis, either in isolation when calculating costs per life years gained or when combined with quality of life experienced in these years to estimate quality-adjusted life years (QALYs). This is the approach favoured by governmental agencies in a number of countries including the United Kingdom (UK), Canada, Australia, the Netherlands, and Sweden (Canadian Agency for Drugs and Technologies in Health, 2006; Claxton et al., 2002; National Institute for Health and Clinical Excellence (NICE), 2013). In this article, we focus on the methodology for estimating the impact on length of life.

As full survival data are rarely available, the evaluation problem faced can be broken down into two key aspects: estimating the effect of the policy on mortality, and evaluating the long-term gains in life years associated with this effect on mortality (Angrist and Pischke, 2008; Heckman, 2008; Jones and Rice, 2011). Policy evaluations attempting to take a lifetime horizon can use administrative data sets to estimate changes in short-term mortality and subsequently convert these to projected gains in life years using published estimates of life expectancy for the general population. Examples include measuring National Health Service (NHS) productivity (Castelli et al., 2007; Dawson et al., 2005), estimating the National Institute for Health and Care Excellence (NICE) decision threshold (Claxton et al., 2015), and analysing the cost-effectiveness of pay-for-performance (P4P) (Chapter 2).

The approach taken in previous work has been to estimate the impact of a programme in terms of changes in the probability of mortality within 30 days, assessed as a binary outcome (Chapter 2; Castelli et al., 2007; Dawson et al., 2005). Estimated reductions in this mortality rate are then translated into life years gained. Patients dying within 30 days are effectively assumed to die instantly and attributed no survival days in this calculation, whilst those surviving past 30 days are assigned the remaining age-gender specific life expectancy of the general population.

These published estimates of life expectancy at particular ages are calculated from mortality rates observed in the general population. Although they appear to be projections, life expectancy figures are in fact a summary statistic of cross-sectional age-specific mortality rates. Life expectancy figures therefore represent the average length of life of a hypothetical cohort of individuals exposed for each of their remaining years to the age-specific annual mortality rates experienced by the general population who were alive at the start of a reference period. Life expectancy is positive at each age, and the implied length of life (years lived so far plus remaining life expectancy) increases with age. Thus, whilst life expectancy at birth is 83 years for females in England, life expectancy for those who survive to age 83 years is 8 years (Office for National Statistics, 2014).

The length of life of the patients affected by health care policies and programmes is, however, likely to differ from that of the general population. This may lead to incorrect estimation of the effects on life years gained as a result of any reductions in mortality rates. The true impact of such programmes upon survival may also be more complex, with changes to health care initiatives having the potential to impact survival over the whole life course. These longer-term effects are not captured in evaluations focusing solely on mortality rates within the short-term windows normally assessed. Even with the minimal data of one financial year available in many administrative data sets, it is possible to observe the majority of patients for longer than the standard period of 30 days, unless they are treated during the last month of the period. This enables observation of these patients for an additional 1 to 334 days, depending on when in the year they entered treatment. This prolonged follow-up information has, however, often been ignored in policy evaluations to date.

When analysing data from clinical trials, survival models are commonly used to extrapolate gains in life expectancy from the observed trial data (Grieve et al., 2013; Latimer, 2013). Such analysis utilises all available follow-up information on patients rather than applying an arbitrary cut off window. In this article, we examine whether the additional information available within administrative data sets on survival beyond the usual 30 days considered, albeit censored, can be used to improve the accuracy of estimated life years gained in policy evaluations. The aim of this article is to demonstrate the feasibility and materiality of using parametric survival models commonly employed in clinical trials analysis to extrapolate future survival for use in health care policy evaluations.

4.2 Methods

We used our previous cost-effectiveness analysis of the first 18 months of the Advancing Quality (AQ) P4P programme as a motivating example (Chapter 2). AQ is a quality improvement initiative, supported by financial incentives, introduced to all of the 24 hospitals in the North West of England in October 2008 (see Chapters 2 and 3 for a full description of the policy). We previously estimated that the introduction of AQ led to a 1.6 percentage point reduction (95% CI = -2.4 to -0.8) in the rate of mortality within 30 days of admission to hospital for pneumonia (Chapter 2). This reduction in mortality was then translated into an estimated gain of 4,701 QALYs by applying published estimates of life expectancy from the general population, which were adjusted for quality of life and discounted.

In our previous analysis, we considered all patients admitted for pneumonia in England over a three year period, including 18 months before the programme was introduced and the first 18 months of its operation. In this article, we consider, for simplicity, a more typical situation in which data on dates of admission and death are available for one financial year prior to the introduction of the programme and one financial year following its implementation.

We use parametric survival models to estimate the effect of AQ on survival amongst the affected population over a lifetime horizon. These results were compared to those obtained by estimating the impact of the policy on mortality 30 days after admission and applying general population life expectancy estimates to this short-term mortality change.

4.2.1 Data

We used individual patient-level data from national Hospital Episode Statistics for patients admitted to hospital in England between 1st April 2007 and 31st March 2008 and between 1st April 2009 and 31st March 2010. These were linked to Office for National Statistics (ONS) death records (Health and Social Care Information Centre, 2013) for the period 1st April 2007 to 31st March 2011, the latest date on which the death records were complete at the time of data extraction.

We restricted the analysis to patients admitted in an emergency with pneumonia using International Classification of Diseases 10th Revision (ICD-10) codes for the rules specified for the AQ scheme[6]. Secondary ICD-10 diagnosis codes were used to identify patients with Elixhauser conditions (Quan et al., 2005), which were used to risk-adjust our estimates in conjunction with information on the primary diagnosis, age, gender, financial quarter of admission, hospital Trust, the location from which a patient was admitted (own home or institution), and the type of admission (emergency or transfer) (Bottle et al., 2014; Gutacker et al., 2015).

4.2.2 Comparison of methods on a development cohort

We first used the cohort of patients admitted to any hospital in England prior to the introduction of AQ (1st April 2007 to 31st March 2008) to compare three methods for estimating the remaining life years of the population using data from this financial year only. We then compared the predicted survival to the observed data on the survival of the cohort up to 31st March 2011.

The purpose of this initial analysis was to illustrate the difference in the magnitude of the estimated remaining life years of a patient population when the additional information available on survival past 30 days is utilised and information on the risk of death is taken from the population under investigation rather than general population figures. In addition, this exercise was used to select the most appropriate functional form for the survival models to be used in the later evaluation of AQ.

4.2.2.1 Simple application of published life expectancy tariffs (method i)

We started by applying a simplified version of the method used in our original analysis of the programme, in which mortality occurring within 30 days of admission is defined as a binary outcome (Chapter 2). This method is simplified here in that it does not incorporate quality of life adjustments or discounting, and closely resembles that applied in other policy evaluations (Castelli et al., 2007; Claxton et al., 2015; Dawson et al., 2005). Gender-specific life expectancy estimates

[6] Primary diagnosis of J13, J14, J15, J16.0, J16.8, J18.0, J18.1, J18.2, J18.8 or J18.9, or a primary diagnosis of A40.0, A40.1, A40.2, A40.3, A40.8, A40.9, A41.0, A41.1, A41.2, A41.3, A41.4, A41.5, A41.8, A41.9, J96.0 or J96.2 with a secondary diagnosis from the list of primary diagnosis

at each single year of age from age 18 to 100 years were taken from the 2008 to 10 interim life tables from the ONS (Office for National Statistics, 2014), and attached to patients surviving beyond this 30 day period to estimate their remaining life expectancy:

$$L_i^{ga} = s_i^{30} \cdot L^{ga} \qquad (8)$$

Where s_i^{30} equals 1 if individual *i* survives more than 30 days and 0 otherwise, and L^{ga} is the life expectancy of an individual of gender *g* who is currently aged *a*.

This method implicitly assumes that individuals surviving beyond 30 days after admission survive, on average, the life expectancy of the general population of the same age and gender. This will produce an inaccurate estimate of the actual life expectancy for two reasons;

1. The period of survival within 30 days is not incorporated into the estimate, and
2. It assumes that the life expectancy of individuals who survive past 30 days after admission will be equal to that of the general population of their age and gender.

Moreover, this method ignores information on observed survival available within the data set beyond the period of 30 days after admission.

4.2.2.2 Short-term observed survival plus application of published life expectancy tariffs (method ii)

We then extended this method to utilise all of the information on mortality available within the year of data (1st April 2007 to 31st March 2008) as we could follow patients for between 1 and 365 days depending on their admission date. For those who died during the period, the number of days survived between the date of admission and the date of death is used. Age- and gender-specific estimates of life expectancy were again applied to all patients who remain alive at the end of the observed data period:

$$L_i^{ga} = s_i^{t*} \cdot L^{ga} + (1 - s_i^{t*}) \cdot (t_i^{\dagger} - t_i^{0}) \qquad (9)$$

Where s_i^{t*} is a binary indicator equal to 1 if individual *i* survives to the end of the observation period t^*, $t_i^{\dagger}$ is the date of death for individuals who die before the end of the observation period, and t_i^0 is the date of admission. This improves on the original method by eliminating problem 1 and reducing, but not eliminating, inaccuracies due to problem 2.

4.2.2.3 Extrapolation using survival models (method iii)

Finally, we improved the method used for extrapolation beyond the observed period by estimating parametric survival models on the observed one year of data. These survival models are then used to predict lifetime survival based on the mortality rates of the population of interest observed during this period.

Six standard parametric models were considered (exponential, Weibull, Gompertz, log-logistic, log-normal, and generalised gamma). The fit of these six different models to the observed data was assessed using the Akaike Information Criterion (AIC), tests of whether restrictions on the parameters in the generalised gamma model suggest it could be reduced to the simpler models that it nests, and examination of residual plots, in accordance with the recommendations made by Latimer (2013). The external validity of the extrapolations produced was then assessed by comparing the proportion of the cohort predicted to be alive at annual intervals to the observed survival now available to 31st March 2011.

The risk-adjustment covariates listed in the data section (primary diagnosis, secondary diagnosis, single year of age interacted with gender, financial quarter of admission, hospital Trust, the location from which a patient was admitted (own home or institution), and the type of admission (emergency or transfer)) were included in all of the survival models in the scale parameter using the ‘streg’ command in Stata. The addition of covariates to the shape parameter(s) for models other than the exponential was explored, but did not improve model fit. The shape parameters in all of the models are therefore estimated directly, whilst the scale parameters are estimated as a linear function of the covariates listed.

Our early investigations showed that while standard parametric models were able to fit the observed data well, the tails of these distributions did not correctly represent the pattern of future mortality. This is because the hazard rates experienced by our patient cohort change over time, with the extremely high-risk period shortly after an emergency hospital admission not representative of the lifetime risk of those surviving past this period.

As a result, we estimated survival in two separate models: one for the short-term and one for the longer-term. Short-term survival during the first year was estimated on the observed one year of data. The extrapolation of long-term survival was based on a model estimated on data excluding the first 30 days following admission (Bagust and Beale, 2013; Grieve et al., 2013). These long-term models represent the hazards experienced by our patient cohort after the initial high-risk period following an emergency hospital admission. These are still much larger than those experienced by the general population, but are significantly lower than when they were first admitted to hospital.

This approach bears some similarities to that suggested by Gelber and colleagues in that survival is divided into the short-term and the tails of the distribution which are fitted separately (Gelber et al., 1993), but here, we fitted a parametric model to the short-term data rather than simply using the observed Kaplan-Meier curve. This allowed us to estimate the effect of covariates on survival in both the observed and extrapolated periods.

Following estimation of the survival models, we created additional rows of data for each individual for each possible future year up to the age of 100 years. We estimated the probability of surviving to that year, allowing for the progression of time and increments in age. This approach is analogous to the estimation of transition probabilities in a Markov model:

$$\widetilde{m}_i^t\ (a_{i0},\, x_i) = \left[\frac{s_{it}\,(a_{it},x_i)}{s_{i,t-1}\,(a_{it},x_i)}\right] - 1 \qquad (10)$$

where $\widetilde{m}_i^t$ is the probability that individual *i* will die by time *t,* given that they have survived to time *t-1,* and s_{it} is the probability that individual *i* will survive to time *t,* given the values of their covariates *x* and their age a_i at the time of their admission. We estimated the probability of dying during the first year ($\widehat{m}^1$) using all data on survival following the admission date (short-term model) and the probability of dying in subsequent years ($\widetilde{m}^t$) using the data on survival following 30 days after the admission date (long-term model).

We then calculated the individual's life expectancy using the sum of the probability of surviving to the end of the first year and the survival rates for each subsequent year up to the maximum age of 100 years:

$$L_i = (1 - \widehat{m}_i^1) \cdot (t^* - t_i^0) + \sum_{j=a_{i0}+1}^{A} s_{i,j-a_{i0}}\left(a_{ij}, x_i\right) \cdot (1 - \frac{1}{\widetilde{m}_i^{j+1-a_{i0}}}) \qquad (11)$$

where L_i is the life expectancy of individual *i,* $\widehat{m}_i^1$ is the probability that individual *i* will die by the end of the first year, t^* - t_i^0 is the length of time between the individual's admission date and the end of the first year, *A* is the maximum age (100 years), and the summation is over products of the probability of surviving to the start of each subsequent year and the probability of surviving that subsequent year, given that the individual will have aged by those subsequent years.

This method again eliminates problem 1 (the period of survival within 30 days not being incorporated into the estimate) and further reduces inaccuracies due to problem 2 (previously assuming that the life expectancy of individuals who survive past 30 days after admission will be equal to that of the general population of their age and gender) by using information on the mortality rates of the population under study to estimate their future survival. We compared the results given at each of these three stages, as the original assumptions were dropped and improved upon, to illustrate the materiality of these developments to our estimates of life years remaining.

4.2.3 Application to the evaluation of Advancing Quality

Having demonstrated the use of parametric survival models, and the materiality of the difference that this makes to the estimated life years remaining for our patient cohort, we illustrate how these models can be used in an applied programme evaluation. We considered a dichotomous difference-in-differences (DiD) design in which outcomes were observed for treated and control units before and after the introduction of the programme:

$$L_{ijt} = f(a + X'b + u_j + v_t + \delta D_j^1 \cdot D_t^2 + \varepsilon_{ijt}) \qquad (12)$$

where L_{ijt} is the life expectancy of individual *i* treated in hospital *j* at time *t*, *f*(·) is the link function, *X* is the vector of case-mix covariates, u_j are provider fixed effects, v_t are time fixed effects, D_j^1 is a dummy variable taking the value 1 for hospitals that become part of the AQ programme, D_t^2 is a dummy variable taking the value 1 in the periods after the introduction of the AQ programme, and ε_{ijt} is an individual-specific error terms. δ is the DiD term, which is our coefficient of interest.

We first considered the situation outlined above in which data on dates of admission and death were available for one financial year prior to the introduction of the programme (1st April 2007 to 31st March 2008) and one financial year following its implementation (1st April 2009 to 31st March 2010). An additional advantage of using survival analysis, however, is that the additional follow-up data on the pre-intervention group collected during the same period as the initial follow-up of the post-intervention group can be utilised. We therefore examined how the life expectancy estimates were affected when including the additional follow-up available (1st April 2008 to 31st March 2010) on the group admitted prior to the intervention, so that this group was now followed up for a maximum of three years. In principle, utilising this additional available information on the pre-intervention population should improve the accuracy of our estimates of long-term survival and the estimated impact of the programme.

To calculate the effect of the programme on life expectancy, we used average partial effects. We estimated life expectancy for the individuals admitted to AQ hospitals in the post-AQ period under two scenarios, with the DiD term set to one and to zero. These represent our estimates of the life expectancy of these patients in the presence and absence of the policy, respectively.

These results were compared to those obtained by estimating the impact of the policy using linear regression on general population life expectancy estimates attached to individuals who survived 30 days after admission. Finally, we performed a sensitivity analysis using the second best-fitting parametric model to estimate the impact of the policy on life expectancy to illustrate the impact of model selection on these estimates.

4.3 Results

4.3.1 Development cohort

The characteristics of the patient cohort are presented in Table 20 and discussed in more detail below when we describe the application of our method to AQ (section 4.3.2). The annual mortality rates by age and gender for this patient cohort were considerably higher than those experienced by the general population (Table 21), illustrating the importance of using information on the risk of death from the population under investigation rather than general population figures when estimating remaining life years. Using general population figures would lead us to underestimate the annual mortality rate experienced by our population by a factor of between 2 (age >100 years) and over 300 (age 20 years). Figure 3 shows the Kaplan-Meier survival curve for these patients over the period from 1st April 2007 to 31st March 2008 and highlights the high rate of mortality in the initial high-risk period following an emergency admission.

Table 20: Descriptive statistics for patients admitted for pneumonia, by region and time period

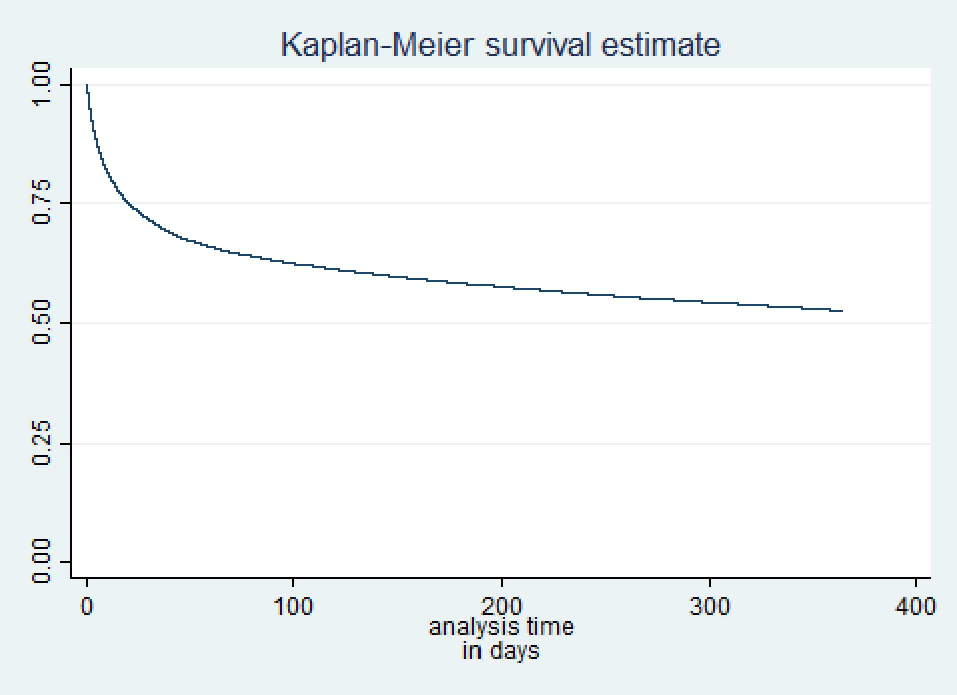
	Patients admitted in 2007/08		**Patients admitted in 2009/10**	
	North West	Rest of England	North West	Rest of England
Admitted patients, n	17,993	95,296	19,946	106,365
Age at admission, years mean (range)	71.7 (18-106)	72.2 (18-107)	71.9 (18-110)	72.8 (18-110)
Female patients, %	49.8%	48.7%	50.3%	49.1%
Mean number of coexisting conditions	1.79	1.65	1.99	1.92
Unadjusted mortality within 30 days, %	28.4%	27.3%	25.6%	26.0%
Dead by the end of the financial year, %	40.7%	38.6%	37.3%	37.3%

Table 21: Annual mortality rates by age and gender for the general population and admitted patients

	Annual mortality rates Females		**Annual mortality rates Males**	
Age (years)	General population, %	Patients admitted for pneumonia in 2007/08, % (n)	General population, %	Patients admitted for pneumonia in 2007/08, % (n)
20	0.02%	6.12% (98)	0.06%	3.60% (111)
30	0.04%	4.71% (191)	0.09%	5.11% (176)
40	0.10%	11.33% (309)	0.17%	8.36% (311)
50	0.24%	17.53% (291)	0.35%	20.15% (402)
60	0.56%	27.23% (584)	0.88%	30.42% (733)
70	1.46%	42.78% (783)	2.24%	46.33% (980)
80	4.52%	59.63% (1,469)	6.51%	66.77% (1,580)
90	14.60%	77.50% (1,142)	16.93%	84.53% (808)
100	39.19%	89.90% (109)	41.79%	100.00% (29)

Notes: Source of general population figures: ONS interim life tables 2008 – 10. Source of pneumonia patient mortality: authors' analysis of Hospital Episode Statistics linked to ONS death records.

Figure 3: Kaplan-Meier survival curve for the cohort of patients admitted for pneumonia in 2007/08



A comparison of the performance of the six parametric survival models showed that the generalised gamma distribution gave the lowest AIC, followed by the log-normal distribution (Table 22). A Wald test confirmed that the generalised gamma does not reduce to a log-normal distribution in this case (p<0.001). Finally, the generalised gamma gave the best performance on the external validity assessment, predicting the proportion of the cohort alive to within 1% of the observed survival rate at each of the four annual time points now available in the prolonged follow-up data (Table 22). A generalised gamma distribution was therefore chosen to model survival.

The life expectancy of the cohort of 113,289 patients admitted during 2007/08 was estimated using the three different methods (Table 23). 27% of the cohort died within 30 days and were therefore assigned no life expectancy under method i. The remaining 73% surviving past this point were assigned life expectancy estimates from the general population. This approach estimated that the cohort had on average 13.15 years of life remaining.

Table 22: Internal and external validity of difference parametric survival functions

	Exponential	**Weibull**	**Gompertz**	**Log-normal**	**Log-logistic**	**Generalized gamma**	
Internal validity							
AIC	326,943.4	288,141.2	291,562.7	283,530.6	285,139.2	283,386.1	
External validity							
Time point	**Predicted survival**						**Observed survival**
31 Mar 2008	56.76%	60.02%	60.02%	60.10%	59.78%	60.21%	61.05%
31 Mar 2009	36.02%	46.51%	52.87%	49.08%	47.63%	48.96%	49.73%
31 Mar 2010	25.93%	39.26%	52.37%	43.92%	41.77%	43.67%	43.86%
31 Mar 2011	20.04%	34.30%	52.24%	40.49%	37.92%	40.14%	39.31%

Table 23: Comparison of estimates of remaining life years for patients admitted for pneumonia 2007/08

Assessment period	**Extrapolation method**	**Number (%) alive at end of assessment period**	**Estimated life years remaining, mean**
Admission to 30 days later	General population life expectancy	82,208 (72.56%)	13.15 (SD 14.65) range 0 – 64.88
Admission to end of financial year	General population life expectancy	69,158 (61.05%)	11.98 (SD 14.98) range 0 – 64.88
Admission to end of financial year	Parametric survival models	69,158 (61.05%)	9.19 (SD 10.94) range 0.001 – 75.17

When taking into account additional information on survival past 30 days to the end of the financial year, a further 12% of the initial cohort was observed to have died during this period. The impact of using this additional available information on survival past 30 days was to reduce the estimate of average life years remaining from 13.15 years to 11.98 years. This estimate, however, still assumes that patients surviving to the end of the financial year under observation will experience, on average, the life expectancy of the general population of their age and gender.

Finally, this assumption was relaxed when we used parametric survival models with a generalised gamma distribution to predict life expectancy based upon the rates of mortality observed within the cohort. This method further reduced our estimate of the average number of life years for the cohort to 9.19 years.

In this development cohort, taking into account the additional information available on survival past 30 days reduced the estimate of average life expectancy by 9%. Once survival models were used to extrapolate future survival past the observed period, the original estimate was reduced by 30%.

4.3.2 Application to the evaluation of Advancing Quality

Of the 113,289 patients admitted for pneumonia between 1st April 2007 and 31st March 2008, 17,993 (16%) were admitted to hospitals that would later participate in the AQ scheme (Table 20). Patients admitted to hospitals in the North West of England before the introduction of AQ were slightly younger than their counterparts in the rest of England (71.7 versus 72.2 years, respectively) and had a higher number of coexisting conditions (1.79 versus 1.65, respectively). The unadjusted mortality rate within 30 days of admission was higher in the North West than in the rest of England (28.4% versus 27.3%, respectively). This difference in mortality persisted in the longer-term, with 40.7% of the cohort admitted to hospitals in the North West having died by the end of the financial year versus 38.6% of those admitted in the rest of England.

During the period 1st April 2009 to 31st March 2010, 19,946 patients were admitted for pneumonia to hospitals participating in the AQ programme, and 106,365 were admitted to hospitals in the rest of England which did not participate in AQ. Patients admitted to hospitals in the North West were again slightly younger than their counterparts in the rest of England (71.9 versus 72.8 years, respectively), with a higher number of coexisting conditions (1.99 versus 1.92, respectively). The unadjusted mortality rates decreased in both regions during our evaluation period, with a greater reduction in the North West than the rest of England. The rate of mortality within 30 days of admission was lower in the North West than the rest of England in this period (25.6% versus 26.0%, respectively), with no difference in the proportion of patients still alive at the end of the financial year (37.3% died in both regions). These figures illustrate the positive effect of the programme upon reducing mortality rates within 30 days of admission found in our previous evaluation (Chapter 2).

Table 24 shows the estimated effect of AQ upon remaining life expectancy for patients admitted in the North West in 2009/10. An ordinary least squares (OLS) DiD regression of the general population life expectancy figures applied to those surviving past 30 days post-admission (method i) estimated that AQ led to an average increase in remaining life expectancy of 0.154 years. The remaining life expectancy of the patient cohort was estimated to have been 13.22 years in the presence of AQ, and would have been 13.06 in the absence of the policy.

An OLS DiD regression of the general population life expectancy figures now applied to those surviving past the end of the financial year (method ii) produced a larger treatment effect estimate of AQ on average life expectancy of 0.221 years. This is despite lower estimates of remaining life expectancy of the patient cohort both in the presence (11.98 years) and absence of AQ (11.76 years). These lower absolute estimates of remaining life expectancy are expected as they account for the additional deaths we are able to observe using the extended follow-up to the end of the financial year. The increase in the estimated effect of AQ upon life expectancy indicated that the policy impacted on survival beyond the 30-day post-admission window usually assessed.

In the parametric survival models utilising mortality information until the end of the financial year, the coefficient on the DiD term was negative and statistically significant. The generalised gamma was parameterised in the accelerated failure time (AFT) metric, meaning that coefficients of < 1 indicate that time passes more slowly and so failure (death) would be expected to occur later as a result of AQ. This estimated effect on the failure time translated into an average estimate of 9.04 remaining life years of the cohort in the presence of the policy, and 8.73 in its absence. These reductions in both estimates of average life expectancy compared to the estimates from the OLS models were again as expected, as we now used information on the mortality rates observed among the patient population rather than general population life expectancy estimates. Nevertheless, this translated into a larger estimated treatment effect of AQ of 0.311 years, suggesting that AQ had a prolonged impact on survival past the end of the financial year assessed.

We then estimated survival models utilising the additional data available on follow-up for the pre-intervention group to the end of our evaluation period (data window from 1st April 2007 to 31st March 2010) to inform our estimates. This additional data increased the precision of our estimates, with the treatment effect of AQ now estimated to be 0.380 years. Utilising this additional data slightly decreased the estimated remaining life expectancy for the cohort both in the presence and absence of the policy (8.439 versus 8.059 years, respectively), but further increased the estimated effect of AQ on the life expectancy of patients admitted to hospitals in the North West during the treatment period.

Finally, we present a sensitivity analysis using the lognormal distribution (Table 25). This analysis produced very similar results to that using the generalised gamma over the same data period. The log-normal distribution was again parameterised in the AFT metric, meaning that coefficients of < 1 are associated with a deceleration of time to death. The remaining life expectancy of the patient cohort admitted to hospitals in the North West after the policy was introduced was estimated to be 9.284 years in the presence of AQ and 8.971 years in its absence. This resulted in an estimated treatment effect attributable to AQ of 0.313 years. In this instance, the choice of distribution used to model survival had little impact on the treatment effect estimates.

Table 24: Estimated effect of Advancing Quality on the remaining life expectancy of patients admitted to hospitals in the North West in 2009/10

	Method i	Method ii	Method iii		Method iii	
Source of life expectancy estimates	General population life tables	General population life tables	Survival analysis using one financial year of follow-up		Survival analysis using all available follow up	
			Short-term model: Entry time = admission	Long-term model: Entry time = 31 days post admission	Short-term model: Entry time = admission	Long-term model: Entry time = 31 days post admission
Estimates						
Estimation method	OLS	OLS	Generalised Gamma	Generalised Gamma	Generalised Gamma	Generalised Gamma
Follow-up periods						
Patients admitted before AQ	t_i^0 to t_i^{30}		t_i^0 to t^1	t_i^{30} to t^1	t_i^0 to t^3	t_i^{30} to t^3
Patients admitted after AQ	t_i^0 to t_i^{30}		t_i^0 to t^3	t_i^{30} to t^3	t_i^0 to t^3	t_i^{30} to t^3
Coefficient on the difference-in-differences term	0.154	0.221	0.103	0.089	0.142	0.101
	(2.39)	(3.04)	(2.64)	(1.71)	(3.62)	(2.26)
Ln(σ)			1.095	0.876	1.105	0.731
			(301.02)	(77.08)	(364.60)	(88.55)
κ			-0.219	0.0639	-0.0587	0.388
			(-16.25)	(2.84)	(-5.42)	(25.43)
Observations	239,600	239,600	239,600	156,860	239,600	164,438
No. of deaths	63,845	91,272	91,272	26,785	110,747	45,290
Extrapolations						
Life expectancy for those admitted in the North West in 2009/10 (years)	13.218	11.982	9.041		8.439	
Counterfactual estimated life expectancy for those admitted in the North West in 2009/10 in the absence of programme (years)	13.064	11.761	8.730		8.059	
Estimated effect of programme on life expectancy for those admitted in the North West in 2009/10 (years)	0.154	0.221	0.311		0.380	

Notes: Robust *t* statistics in parentheses. Only coefficients of interest are shown. Models also control for primary and secondary diagnoses, age, gender, financial quarter of admission, hospital trust, the location from which a patient was admitted (own home or institution), and the type of admission (emergency or transfer). t_i^0 is the date of admission for individual *i*; t_i^{30} is 30 days after the date of admission for individual *i; t^1* is a 1 year follow up period (to 31st March 2008) *; t^3* is a 3 year follow up period (to 31st March 2010).

Table 25: Sensitivity analysis using the log-normal distribution to estimate the effect of Advancing Quality on the remaining life expectancy of patients admitted to hospitals in the North West in 2009/10

	Method iii	
Source of life expectancy estimates	Survival analysis using one financial year of follow-up	
	Short-term model: Entry time = admission	Long-term model: Entry time = 31 days post admission
Estimates		
Estimation method	Log-normal	Log-normal
Follow-up periods		
Patients admitted before AQ		
Patients admitted after AQ		
Coefficient on the difference-in-differences term	0.106 (2.70)	0.089 (1.71)
Ln(σ)	1.048 (470.06)	0.903 (192.77)
Observations	239,600	156,860
No. of deaths	91,272	26,785
Extrapolations		
Life expectancy for those admitted in the North West in 2009/10 (years)	9.284	
Counterfactual estimated life expectancy for those admitted in the North West in 2009/10 in the absence of programme (years)	8.971	
Estimated effect of programme on life expectancy for those admitted in the North West in 2009/10 (years)	0.313	

Notes: Robust *t* statistics in parentheses. Only coefficients of interest are shown. Models also control for primary and secondary diagnoses, age, gender, financial quarter of admission, hospital trust, the location from which a patient was admitted (own home or institution), and the type of admission (emergency or transfer).

4.4 Discussion

Policy evaluations attempting to take a lifetime horizon have used administrative data sets to estimate changes in short-term mortality and subsequently converted these to projected gains in life years using published estimates of life expectancy for the general population. This may lead to inaccurate estimates of the effect on life years gained if the length of life of patients affected by the health care programme differs from that of the general population or if the policy affects survival over the whole life course rather than just during the evaluation period.

While statistics such as mortality occurring within 30 days of admission are a useful indicator of a programme’s success, information on the impact over the lifetime horizon of the affected patients is needed to inform decisions regarding cost-effectiveness. We have demonstrated the feasibility of using parametric survival models commonly employed in clinical trials analysis to extrapolate future survival in health care policy evaluations. Our application of these methods to the AQ initiative reinforces the importance of both using the mortality rates observed in the patient population under study rather than taking estimates from the general population and utilising all available follow-up data on survival.

We have demonstrated the impact that this has on the estimates of both the remaining life years of a patient cohort and the treatment effect of the policy under investigation. In a cohort of patients admitted to hospital for pneumonia during one financial year, the estimated mean life years remaining was 30% lower when parametric survival models were used compared to the traditional method of applying general population life expectancy estimates to those surviving more than 30 days past admission. When assessing the predictive accuracy of our chosen survival model against the further three years of follow-up data now available, predictions of the proportion of the cohort alive at four annual intervals were within 1% of the observed survival rate, supporting the accuracy of this method.

However, when survival analysis was used to estimate the effect of the AQ programme on the survival of the patients treated, this produced a larger estimated treatment effect than the traditional method. This suggests that AQ impacted on survival past the 30 day post-admission window usually assessed. The ability of survival models to capture the effect of a policy over the whole life course of the affected patients is another advantage of using this method. Nevertheless, we acknowledge that even extrapolations based on extended follow-up may still provide inaccurate predictions, as illustrated by Davies and colleagues (2013). A conservative approach to extrapolation is therefore recommended, fitting and testing a range of survival models to assess both their internal and external validity. While overcoming the assumption of general-population life expectancy, parametric modelling introduces new assumptions that must be considered.

In addition to following the useful guidance on survival analysis published by Latimer (2013), there are some further considerations for researchers performing programme evaluations using administrative data rather than health technology assessments (HTAs) of single interventions from randomised controlled trials. A first step should always be a comparison of the observed mortality rates of the patient population under consideration to those experienced by the general population. If the mortality risk is only apparent in the very short-term, and longer-term survival rates are similar to those experienced by the general population, then survival analysis may not be needed.

If, as in the case of our evaluation, the mortality rates of the patient population are significantly different to the general population, various survival models should be assessed. In a pre- and post-evaluation design, survival models can be developed on the pre-intervention population and their predictive performance evaluated against the observed follow-up available on these patients during the post-intervention period to assess the external validity of the models developed. All available follow-up data on the pre-intervention group can then be utilised to inform the baseline pattern of survival and increase the accuracy of the estimated treatment effect of the programme in question. If modelling short- and long-term survival separately as done here, a simple histogram of events over time can be informative in selecting the cut-off point at which data is excluded when fitting the long-term models to extrapolate. The trade-off between the loss of follow-up data utilised and the exclusion of short-term event rates not representative of long-term survival must be considered, and will largely depend on the length of follow-up available.

4.4.1 Limitations and directions for future work

The purpose of this article is to demonstrate how survival analysis can be applied beyond the setting of randomised controlled trials in order to extrapolate survival for use in cost-effectiveness analysis of health care policies and programmes using administrative data sets. In order for the estimates of life years gained calculated here to be used in a cost-effectiveness analysis, the stream of remaining life years for each patient under evaluation would also need to be adjusted for quality of life and discounted to present day values in order to calculate the effect of AQ in terms of QALYs. These extensions are simple to perform, as demonstrated in our previous evaluation (Chapter 2).

The large scale of administrative data sets such as that used here could also offer a useful source for researchers wishing to develop and refine further methodological advances in survival analysis. Unlike randomised controlled trials, administrative data can often allow researchers to capture the entire affected population of interest in a real life treatment setting and so enable an externally valid evaluation of the effect of the policy or programme change in question. The scale of these data sets and amount of additional information that they contain have the potential to enable accurate estimates of survival using minimal follow-up.

We found that previously employed methods used to estimate the impact of health care policies over a lifetime horizon led us to overestimate the remaining life expectancy of our cohort but underestimate the impact of the policy in question on this life expectancy. The application of survival analysis utilising minimal additional, but readily available, data on prolonged follow-up considerably improved the accuracy of our estimates of both absolute survival and the impact of the AQ programme on the survival of the targeted patient population. There are many national administrative data sets available for use in similar analysis such as Hospital Episode Statistics in England (Health and Social Care Information Centre, 2012), insurance claims databases in America, and linked administrative datasets in Canada (Population Data BC, 2016). We hope that the methods demonstrated here will be further developed and applied using these data sets to improve the accuracy of future cost-effectiveness analyses of health care policies and programmes.

5. What are the costs and benefits of providing comprehensive seven-day services for emergency hospital admissions?

Abstract

The English National Health Service is moving towards providing comprehensive seven-day hospital services in response to higher death rates for emergency weekend admissions. Using Hospital Episode Statistics between 1st April 2010 and 31st March 2011 linked to all-cause mortality within 30 days of admission, we estimate the number of excess deaths and the loss in quality-adjusted life years associated with emergency weekend admissions. The crude 30-day mortality rate was 3.70% for weekday admissions and 4.05% for weekend admissions. The excess weekend death rate equates to 4,355 (risk adjusted 5,353) additional deaths each year. The health gain of avoiding these deaths would be 29,727 to 36,539 quality-adjusted life years per year. The estimated cost of implementing seven-day services is £1.07bn to £1.43bn, which exceeds by £339m to £831m the maximum spend based on the National Institute for Health and Care Excellence threshold of £595m to £731m. There is as yet no clear evidence that seven-day services will reduce weekend deaths or can be achieved without increasing weekday deaths. The planned cost of implementing seven-day services greatly exceeds the maximum amount that the National Health Service should spend on eradicating the weekend effect based on current evidence. Policy makers and service providers should focus on identifying specific service extensions for which cost-effectiveness can be demonstrated.

This chapter is published in *Health Economics.*

Citation: Meacock R, Doran T, Sutton M. (2015). What are the costs and benefits of providing comprehensive seven-day services for emergency hospital admissions? *Health Economics,* 24(8), 907-912.

5.1 Introduction

There is growing evidence that patients admitted to hospital in an emergency outside of normal working hours, when staffing levels and access to ancillary, diagnostic and support services are lower and senior clinical staff are located away from the premises, have a higher risk of death (NHS England, Seven Days a Week Forum, 2013a). The National Health Service (NHS) was founded on the principle of equitable care, and variation in outcomes by time of admission therefore represents a failure of the service to meet one of its most fundamental obligations.

The NHS in England has responded with a commitment to make routine services available seven days a week (NHS Commissioning Board, 2013), and the Academy of Medical Royal Colleges has recommended that hospital inpatients should be reviewed by an on-site specialist every day, including at weekends (Academy of Medical Royal Colleges, 2012). The latest planning guidance for the NHS requires acute care providers to implement at least five of ten clinical standards for seven-day services in the 2015/16 financial year (NHS England, 2014), and commissioners have been encouraged to use local financial sanctions under the Commissioning for Quality and Innovation (CQUIN) (Kristensen et al., 2013) framework to ensure that progress is made.

However, providing the same level of services every day of the week may not be the most cost-effective way of distributing limited healthcare resources. It is not yet known whether changing to a seven-day service will improve outcomes, and what the costs of any such re-organisation will be. In this paper, we discuss the evidence base being used to support the case for seven-day services and, using national statistics and the results of published studies, we estimate the potential benefits of introducing such service extensions across England compared to the costs of doing so.

5.2 The evidence on seven-day services

5.2.1 Evidence for increased risk at weekends

Recent evidence from England suggests that patients' risk-adjusted probability of dying is increased by 11% (95% CI: 9 to 13%) if admitted to hospital on a Saturday, and 16% (95% CI: 14 to 18%) if admitted on a Sunday, compared to those admitted mid-week (Freemantle et al., 2012). This 'weekend effect' varies substantially by condition, from a zero additional risk for pneumonia to 16% for stroke, 28% for lung cancer, and 37% for renal failure (Freemantle et al., 2012). To interpret these weekend effects, it is important to consider the baseline level of risk. In the review commissioned by the NHS Seven Days a Week Forum (NHS England, Seven Days a Week Forum, 2013b), only one study by Aylin and colleagues (2010) presented actual mortality rates for patients admitted at the weekend and during the week. Using data from 2005/06, Aylin et al. reported in-hospital mortality rates for emergency admissions of 4.9% for weekday admissions and 5.2% for weekend admissions (Aylin et al., 2010). This represents a 10% increase in relative risk, but only a 0.3 percentage point increase in absolute risk.

5.2.2 Evidence on consultant cover at weekends

One of the main differences between hospital care at weekends compared with weekdays is the reduced availability of senior clinical staff, and this is often cited as an explanation for the observed weekend effect (Bell and Redelmeier, 2001; NHS England, Seven Days a Week Forum, 2013a; Schmulewitz et al., 2005). However, there is a lack of evidence that increasing the levels of consultant cover at weekends leads to reductions in mortality rates (NHS England, Seven Days a Week Forum, 2013b).

Available evidence on the impact of extending service provision comes from a small number of case studies of specific care pathways. For stroke, for example, improved outcomes, reduced length-of-stay and favourable evidence on cost-effectiveness have been found in specialised units configured to treat patients admitted for this condition every day of the week in London (Hunter et al., 2013), although later work showed that the same results were not seen in another location, Greater Manchester (Morris et al., 2014). The death rate within seven days of admission is reported to have fallen from 10.0% to 7.3% for those admitted at the weekend after the reorganisation of stroke services in London (NHS England, Seven Days a Week Forum, 2013b). However, these facilities provide a range of enhanced facilities in addition to seven-day services, and it is therefore difficult to identify which aspects were responsible for the observed improvements in outcomes. In the London case study, the mortality rate for patients admitted during the week was reported to have fallen from 8.0% to 6.4% (NHS England, Seven Days a Week Forum, 2013b), meaning that the relative weekend effect was reduced but not eliminated.

5.2.3 Cost-effectiveness of extending normal operational hours

Funding new interventions imposes costs on health systems, reducing the resources available for existing services and potentially resulting in a net loss of health benefits (Claxton et al., 2015, 2013). The costs of extending normal operational hours must therefore be weighed against the predicted benefits.

Increasing the level of consultant cover during the weekends will require a redistribution of the existing workforce and/or additional training and recruitment. Diverting consultant cover away from weekdays towards weekends would be expected to affect the quality of services and outcomes for patients admitted during the week. The introduction of seven-day services might therefore narrow the gap between weekday and weekend mortality, but at the cost of higher weekday rates.

5.3 The potential benefits of implementing seven-day hospital services

5.3.1 Methods

We estimated the loss in patient health associated with the weekend effect for emergency admissions to all hospitals in England between 1st April 2010 and 31st March 2011. We used data on inpatient episodes from Hospital Episode Statistics linked to data from the Office for National Statistics (ONS) on all-cause mortality (both in and out of hospital) within 30 days of admission

(Health and Social Care Information Centre, 2013). We selected only the emergency admissions using the admission method field in Hospital Episode Statistics.

We first estimated the number of excess deaths occurring amongst patients admitted at the weekend by applying the crude mortality rate observed for weekday admissions to the volume of patients admitted during the weekend, and subtracted this number of expected deaths from the number observed amongst weekend admissions.

As weekend admissions may represent a different case-mix of patients, we then used risk-adjusted mortality rates. We used the risk-adjusted figures reported in the published studies that have been cited as support for the seven-day services initiative. We applied the inverse of the risk-adjusted odds-ratios reported by Freemantle et al. (2012) and Aylin at al. (2010) to the odds (p/(1-p)) of mortality at weekends observed in our data. This represents the expected odds of mortality if weekend patients experienced the same death rate as those admitted during the week once we control for their risk characteristics. We calculated the risk-adjusted number of excess deaths by multiplying the volume of weekend admissions by the risk-adjusted expected mortality rate, and subtracting this number from the observed number of weekend deaths.

We then used a previously developed methodology to calculate the number of quality-adjusted life years (QALYs) that could potentially be gained if the weekend effect were to be eradicated and these excess deaths were averted. This involved applying a discounted and quality-adjusted life expectancy (DANQALE) tariff to the mortality records, which we developed for our evaluation of the Advancing Quality pay-for-performance programme and is explained in detail in Chapter 2. Using these estimated discounted QALY gains, we calculated the maximum amount that the NHS should be prepared to spend on averting these deaths using the standard threshold of £20,000 per QALY used by the National Institute for Health and Care Excellence (NICE) when assessing whether new health technologies are cost-effective.

5.3.2 Results

The crude 30-day mortality rate was 3.70% for patients admitted during the week and 4.05% for those admitted during the weekend, resulting in an excess death rate of 0.35 percentage points (Table 26). If the crude mortality rate observed during the week applied to patients admitted during the weekend, this would translate into an annual estimate of 4,355 excess deaths occurring nationally at weekends (Table 27). After applying the risk-adjusted odds ratio from Freemantle et al (2012), this figure rose to an estimated 5,353 excess deaths. The risk-adjusted odds ratio reported in Aylin et al (2010) is very similar to that obtained using our crude mortality rates and produces a similar estimate of excess weekend deaths of 4,376. Depending upon the figures used, this translates into a potential health gain of between 29,727 and 36,539 QALYs per year if all excess deaths were to be averted. Using the NICE threshold, the NHS should spend no more than £595m to £731m to achieve a health gain of this size. Using the upper bound of the 95% confidence interval for the odds-ratio reported by Freemantle et al (2012) of 1.145 increases these gains by 14%.

The aforementioned calculations represent the maximum possible gains from introducing seven-day services for three reasons. First, they represent the number of deaths that would be averted if the weekend effect were to be completely eradicated by extending services. Second, the methodology likely over-estimates the potential QALY gains from an averted death, as it assumes that those surviving would enjoy the same quality of life and life expectancy as the general population, conditional upon their age and sex (Chapter 2). Third, our calculations represent the best possible scenario, where benefits to patients admitted at the weekend are achieved without any detrimental effect on outcomes for those admitted during the week.

Table 26: Number of emergency admissions, age, and mortality rates by day of admission, England 1st April 2010 to 31st March 2011

Day of admission	**Mean age of patients admitted, (years)**	**Mean age of patients dying within 30 days of admission, (years)**	**Total number of admissions**	**Crude 30-day mortality rate, (%)**
Monday	50.5	77.1	816,742	3.79%
Tuesday	50.8	77.4	793,807	3.68%
Wednesday	50.8	77.3	777,685	3.65%
Thursday	51.0	77.3	792,822	3.65%
Friday	51.5	77.4	798,866	3.73%
Saturday	50.8	80.0	618,666	4.01%
Sunday	49.8	77.6	614,385	4.09%
Weekday	**50.9**	**77.3**	**3,979,922**	**3.70%**
Weekend	**50.3**	**77.9**	**1,233,051**	**4.05%**

Source: Authors’ analysis of Hospital Episode Statistics linked to Office for National Statistics mortality records

Table 27: Estimates of the excess deaths and QALYs associated with weekend admissions

	Death rate [a]			Excess deaths		
Source of estimates	**Weekday**	**Weekend**	**Odds ratio**	**Number[c]**	**QALYs[d]**	**Maximum spend[e]**
Authors' analysis of Hospital Episode Statistics 2010/11	3.7%	4.0%	1.099 (crude)	4,355	29,727	£595m
Freemantle et al. (2012)	Not reported	Not reported	1.125[b] (risk-adjusted)	5,353	36,539	£731m
Aylin et al. (2010)	4.9%	5.2%	1.100 (risk-adjusted)	4,376	29,870	£597m

QALYs, quality-adjusted life years

[a] Crude death rate

[b] 1.125 is the average of the odds ratios presented separately by day in Freemantle et al. (2012), Saturday = 1.11, Sunday = 1.14

[c] Excess deaths are the number of deaths amongst patients admitted at the weekend minus the number of deaths expected if the risk of mortality estimates for patients admitted during the week applied to patients admitted at the weekend

[d] QALYs associated with excess deaths are the number of QALYs that would be gained if all excess deaths were averted

[e] Maximum amount that the NHS should be prepared to spend on averting these deaths using the standard threshold of £20,000 per QALY used by NICE

5.4 The costs of providing seven-day hospital services

Whilst the potential benefits of extending services appear large, they must be compared to the additional costs of doing so. The NHS Seven Days a Week Forum estimated these costs for eight "successful Foundation Trusts with an interest in seven-day services" (NHS England, Seven Days a Week Forum, 2013c, p.25). The costs were estimated using a costing template and interviews with finance staff, managers, and clinicians, followed by two workshops to agree on methodology and overall findings. The cost estimates were highly variable across the Trusts and included some cost savings associated with reduced length-of-stay and reduced readmissions where these were identified by the Trusts. Caution was emphasised in generalising the results, but overall, it was estimated that the costs of implementing seven-day services would be 1.5% to 2% of total hospital income, equivalent to a 5% to 6% increase in the cost of emergency admissions.

According to Department of Health accounts, national expenditure on hospitals was £71.3bn in the financial year 2013/14 (Department of Health, 2014). Application of the NHS England (2013c) estimates suggests that implementing seven-day services would cost between £1.07bn and £1.43bn. This cost exceeds our estimates of the maximum amount that the NHS should spend to eradicate the weekend effect by a factor of 1.5 to 2.4, or between £339m and £831m.

5.5 Conclusions

Recent initiatives to extend normal hours of hospital operation and to provide more comprehensive seven-day services have been implemented in response to alarming statistics on the gap in mortality rates between patients admitted at the weekend compared with those admitted on weekdays. These statistics, however, are insufficient by themselves to justify a policy change towards extending normal hours of operation into the weekend. There is as yet no clear evidence: that seven-day working will, in isolation, reduce the weekend death rate; that lower weekend mortality rates can be achieved without increasing weekday death rates; or that such reorganisation is cost-effective.

Our analysis indicates that the estimated cost of implementing seven-day services exceeds the maximum amount that NICE would recommend the NHS should be prepared to spend on eradicating the observed weekend effect. A comprehensive roll-out of seven-day services across the NHS is therefore unlikely to be a cost-effective use of resources, particularly as our estimates of potential health benefit represent the upper limit of what is achievable. Given the lack of evidence supporting the impact of service extensions on patient outcomes, the benefits actually realised would likely to be much lower. Furthermore, the consequences for patients admitted during the week also need to be considered, as care for these patients may deteriorate if resources are redistributed.

More – and more nuanced – evidence is required before a policy of providing full seven-day services can be supported. For example, our analysis only considered mortality and associated QALYs as an outcome, which are increasingly recognised as limited measures of outcomes

(Coast, 2004). There may be other detrimental effects on quality and outcomes for patients admitted at the weekend that improved weekend services could address. Whilst the policy debate to date has focused on the excess mortality rates observed for patients admitted in an emergency to hospitals during the weekend, there are likely to be wider consequences, such as the impact on elective activity currently undertaken during the week, and the impact on primary and community services that are also limited at weekends. It is possible that selected service extensions – for specific specialties and at certain times of day – could prove to be cost-effective, but substantial commitments of NHS resources should not be made until these can be identified and robust evidence provided.

6. Higher mortality rates amongst emergency patients admitted to hospital at weekends reflect a lower probability of admission

Abstract

Objective: Patients admitted as emergencies to hospitals at the weekend have higher death rates than patients admitted on weekdays. This may be because the restricted service availability at weekends leads to selection of patients with greater average severity of illness. We examined volumes and rates of hospital admissions and deaths across the week for patients presenting to emergency services through two routes: (a) hospital Accident and Emergency departments, which are open throughout the week; and (b) services in the community, for which availability is more restricted at weekends.

Methods: Retrospective observational study of all 140 non-specialist acute hospital Trusts in England analysing 12,670,788 Accident and Emergency attendances and 4,656,586 emergency admissions (940,859 direct admissions from primary care and 3,715,727 admissions through Accident and Emergency) between April 2013 and February 2014. Emergency attendances and admissions to hospital and deaths in any hospital within 30 days of attendance or admission were compared for weekdays and weekends.

Results: Similar numbers of patients attended Accident and Emergency departments on weekends and weekdays. There were similar numbers of deaths amongst patients attending Accident and Emergency on weekend days compared with weekdays (378.0 versus 388.3). Attending an Accident and Emergency department at the weekend was not associated with a significantly higher probability of death (risk-adjusted odds ratio 1.010).

Proportionately fewer patients who attended Accident and Emergency departments at weekends were admitted to hospital (27.5% versus 30.0%) and it is only amongst the subset of patients attending Accident and Emergency departments who were selected for admission to hospital that the probability of dying was significantly higher at the weekend (risk-adjusted odds ratio 1.054).

The average volume of direct admissions from services in the community was 61% lower on weekend days compared to weekdays (1,317 versus 3,404). There were fewer deaths following direct admission on weekend days than weekdays (35.9 versus 80.8). The mortality rate was significantly higher at weekends amongst direct admissions (risk-adjusted odds ratio 1.212) due to the proportionately greater reduction in admissions relative to deaths.

Conclusions: There are fewer deaths following hospital admission at weekends. Higher mortality rates at weekends are found only amongst the subset of patients who are admitted. The reduced availability of primary care services and the higher Accident and Emergency department admission

threshold at weekends mean fewer and sicker patients are admitted at weekends than during the week. Extending services in hospitals and in the community at weekends may increase the number of emergency admissions and therefore lower mortality rates, but may not reduce the absolute number of deaths.

This chapter is published in *Journal of Health Services Research and Policy.*

Citation: Meacock R, Anselmi L, Kristensen S, Doran T, Sutton M. (2017). Higher mortality rates amongst emergency patients admitted to hospital at weekends reflect a lower probability of admission. *Journal of Health Services Research and Policy*, 22(1), 12-19.

6.1 Introduction

The finding that patients admitted to hospital in an emergency at the weekend have a higher mortality rate than those admitted during the week is well documented (Aylin et al., 2010; Bell and Redelmeier, 2001; Freemantle et al., 2015, 2012; Lilford and Chen, 2015; NHS England, Seven Days a Week Forum, 2013a). However, the cause of this 'weekend effect' is not known. The phenomenon has been attributed to reduced availability of senior clinical staff and reduced access to investigative services in hospitals at weekends (Bell and Redelmeier, 2001; NHS England, Seven Days a Week Forum, 2013a; Schmulewitz et al., 2005), but there is no causal evidence establishing this link (Chapter 5; Aylin, 2015; Crump, 2015; McKee, 2015; NHS England, Seven Days a Week Forum, 2013b). Nevertheless, the existing evidence has been used to support moves by the National Health Service (NHS) in England towards seven-day working (Department of Health, 2015; NHS England, Seven Days a Week Forum, 2013a).

This leap from the detection of a statistical association to a reorganisation of the way in which the NHS is provided and staffed has come under unprecedented criticism (Chapter 5; Aylin, 2015; Godlee, 2015; McCartney, 2015; McKee, 2015). Numerous commentaries have raised serious concerns over the interpretation of the papers that have been used to underpin these service changes, highlighting various alternative explanations for the finding of increased mortality rates amongst those admitted to hospital at weekends.

A major concern is that differences in the severity of patients admitted to hospital at the weekend compared to during the week may not be captured fully by the case-mix variables available in administrative data sets (Aylin, 2015; McCartney, 2015; McKee, 2015). The number of patients admitted to hospital in an emergency is markedly reduced at weekends (Chapter 5; Aylin et al., 2010; Freemantle et al., 2015). This may be because the population is less likely to seek emergency care, accident and emergency (A&E) departments are less likely to admit patients, and/or the limited availability of services in the community at weekends leads to fewer direct admissions to hospital. Higher death rates among the smaller number of patients who are admitted at weekends might partly reflect a higher average severity of illness amongst those who are admitted rather than excess avoidable deaths caused by poorer quality of care on admission.

Better understanding of how patients end up in hospital on different days of the week is required if we are to determine whether the weekend effect is a matter for policy concern or a statistical artefact (McCartney, 2015). Our aim is to investigate whether the weekend effect in mortality amongst admitted patients reflects admission of fewer or sicker patients who are at greater risk of dying. We analyse the variation by day of the week in the volume of admissions and subsequent mortality, stratifying patients by their route of access to hospital. We exploit previously under-utilised data on A&E attendances to investigate whether higher mortality amongst the population of patients admitted to hospital reflects a more stringent admission threshold. We then examine the extent to which the limited availability of services in the community at weekends leads to fewer direct admissions and whether there is a higher mortality rate amongst the restricted number of patients who are admitted via this route.

6.2 Methods

6.2.1 Data

We used individual patient-level data on 12,670,788 A&E attendances and 4,656,586 emergency admissions from Hospital Episode Statistics between 1st April 2013 and 28th February 2014 (Health and Social Care Information Centre, 2012). We used an 11-month study period as data were available for 1st April 2013 to 31st March 2014 and each patient was followed for 30 days after attendance or admission to analyse mortality within this subsequent period. We focused on attendances at Type 1 units, which are consultant-led, multi-specialty 24-hour services with full resuscitation facilities and designated accommodation for the reception of A&E patients. These units exclude single specialty centres, minor injury units and walk-in centres, and account for 99% of emergency admissions via A&E (House of Commons Library, 2015).

The attendance records contain information on the patient's age, gender, ethnic group, diagnosis, arrival by ambulance or other mode, whether the attendance is a first or follow-up visit, where the incident occurred (home, work, educational establishment, or other public place), the type of accident (including road traffic accident, assault, deliberate self-harm, sports injury), whether the attendance was patient-initiated or recommended by a professional in another organisation, the date of attendance, and whether the patient was admitted, discharged, or died in the A&E department.

The admission records contain information on the patient's age, gender, ethnic group, primary and secondary diagnoses classified using International Classification of Diseases 10th Revision (ICD-10), whether the patient was admitted from home or another institution, the date of admission, and whether the patient was admitted via A&E or directly by a general practitioner (GP), through a bed bureau, or by a consultant in a scheduled ambulatory clinic. Each record also contains the date of death if the patient died in hospital.

We analysed attendance and admission records from all 140 non-specialist acute Trusts in England. We linked these records using an encrypted patient identifier to the dates of death of all patients who had died in any hospital in England between 1st April 2013 and 31st March 2014. We focused on deaths within 30 days of attendance or admission.

To control for deprivation, we attached the Index of Multiple Deprivation 2010 score to the attendance and admission records using the patient's lower-layer super output area (LSOA) of residence (Department for Communities and Local Government, 2011). England is divided into 32,844 LSOAs, with a mean population of 1,500 (Office for National Statistics, 2011b, 2011c). We included all records for patients whose area of residence in England was known, excluding 772 A&E attendances (0.006% of records) for which risk-adjustment variables were missing.

6.2.2 Statistical analysis

Throughout our analysis we separate patients by their route of admission to hospital, examining two distinct groups. The first group we examine are patients who access emergency services through A&E, which make up the majority of emergency admissions (House of Commons Library, 2015; National Audit Office, 2013). This includes patients directed by their GP to attend. To examine the importance of selection effects amongst the admitted population due to variations in clinical decisions to admit, we focus initially on the entire population of patients who attend A&E and then restrict the analysis to the subset who are selected for admission.

The second group consists of patients admitted directly to hospital in an emergency by GPs (circumventing the A&E department), through a bed bureau, or by specialists in ambulatory clinics, termed ‘direct admissions’ (Hospital Episode Statistics admission method codes 22, 23 and 24). The availability of these community services is more limited at the weekends compared to during the week and we examine whether this leads to fewer direct admissions and whether there is a higher mortality rate amongst the restricted number of patients who are admitted via this route.

Within these groups we compared the mean numbers of A&E attendances, emergency admissions and deaths per day between each day of the week and between weekdays and weekend days using t-tests.

We used logistic regression to estimate the risk-adjusted probability of dying within 30 days for the entire population of patients attending A&E departments by day of the week. We then estimated the risk-adjusted probability of being admitted to hospital and the risk-adjusted probability of dying for the subset of patients who are selected for admission. The case-mix adjustment in these models included information taken from the A&E attendance records on age, gender, ethnicity, diagnosis, arrival mode, first or follow-up visit, incident location, accident type, referral source, deprivation quintile, month, and hospital attended.

We also used logistic regression to estimate the risk-adjusted probability of dying within 30 days of direct admission by day of the week. The case-mix adjustment in these models included information taken from the admission records on age, gender, ethnicity, primary diagnosis (SHMI-grouped Clinical Classifications Software category), Elixhauser (comorbidity) conditions, admission method, admission source, deprivation quintile, month, and admitting hospital (Bottle et al., 2014; Gutacker et al., 2015; Quan et al., 2005). SHMI-grouped Clinical Classifications Software is a tool for grouping patients into a manageable number of clinically meaningful categories using ICD-10 diagnosis codes (HCUP, 2009; Health and Social Care Information Centre, 2015).

We compared each day to Wednesday and then estimated another model comparing weekend admissions to weekday admissions.

The analysis was undertaken using Stata version 13. We clustered the error terms to account for the multiple observations of some individuals using the Stata command ‘robust cluster(id)’ and summarised the goodness-of-fit of the models using the C-statistic.

6.3 Results

6.3.1 A&E department attendances

The average number of people attending A&E is highest on Monday and lowest on Friday (Table 28). Average numbers of attendances on weekend days are similar to week days.

The characteristics of patients attending A&E on weekdays and weekends are given in Table 29. A slightly higher proportion of patients attending A&E at the weekend are children or younger adults, but similar proportions are in the oldest age groups (90+) on weekend days and week days. Proportions of patients with the most common presentations are similar on weekends and weekdays.

The average number of patients attending A&E on weekend days and dying within 30 days is similar to weekdays (Table 28). The crude death rate following an A&E attendance is significantly lower at the weekend compared to during the week (0.99% versus 1.03%).

The risk adjustment model was strongly predictive of mortality (C-statistic of 0.92). After adjusting for risk, attending A&E at the weekend is not associated with a significantly higher probability of mortality than attending during the week (Table 28). Examining the results for each day separately, attending A&E is associated with small but statistically significant higher probabilities of mortality for Sundays and Mondays compared to Wednesday attendance. These increases in relative risk equate to absolute increases in the risk of death of 0.034 percentage points on Monday and 0.037 percentage points on Sunday, from a baseline of 1.02% on a Wednesday.

Table 28: Accident and Emergency (A&E) department attendances and mortality within 30 days

	Average volume of A&E attendances per day on this day of the week	**Average number of deaths within 30 days following A&E attendance per day on this day of the week**	**Crude mortality rate within 30 days following A&E attendance on this day of the week**	**Risk-adjusted mortality rate within 30 days following A&E attendance on this day of the week**[a] Odds-ratio
Monday	41,416.8	402.9	0.97%	1.034 [1.014, 1.055]
Tuesday	37,470.6	388.1	1.04%	0.994 [0.974, 1.014]
Wednesday	36,932.9	375.6	1.02%	Reference
Thursday	36,815.2	385.6	1.05%	1.010 [0.989, 1.030]
Friday	36,425.6	389.4	1.07%	0.996 [0.976, 1.016]
Saturday	37,165.9	374.9	1.01%	0.997 [0.976, 1.017]
Sunday	39,341.8	381.1	0.97%	1.037 [1.016, 1.058]
Weekday	37,812.2	388.3	1.03%	Reference
Weekend	38,253.8	378.0	0.99%	1.010 [0.997, 1.022]
Difference (Weekend –Weekday)	441.6 [-147.5,1030.8]	-10.3 [-22.3, 1.8]	-0.04% [-0.076%, -0.001%]	
Ratio (Weekend: Weekday)	1.01	0.97	0.96	

[a] Logistic regression models including controls for age, gender, ethnicity, diagnosis, arrival mode, first or follow-up visit, incident location, accident type, referral source, deprivation quintile, month, and hospital attended.

Table 29: Descriptive statistics for patients attending Accident and Emergency (A&E) departments

Variable	Weekday n	Weekend n	Weekday %	Weekend %
	9,074,928	3,595,860	100.00	100.00
Female	4,518,407	1,786,423	49.79	49.68
Age category				
<1 year	234,133	103,920	2.58	2.89
1-4 years	613,465	283,354	6.76	7.88
5-9 years	387,499	163,612	4.27	4.55
10-14 years	464,636	151,745	5.12	4.22
15-19 years	572,628	239,844	6.31	6.67
20-24 years	735,069	321,829	8.1	8.95
25-29 years	676,990	278,679	7.46	7.75
30-34 years	589,870	235,169	6.5	6.54
35-39 years	491,861	192,738	5.42	5.36
40-44 years	524,531	201,368	5.78	5.6
45-49 years	525,438	199,211	5.79	5.54
50-54 years	477,341	178,714	5.26	4.97
55-59 years	393,852	146,352	4.34	4.07
60-64 years	362,090	135,924	3.99	3.78
65-69 years	378,424	142,396	4.17	3.96
70-74 years	347,570	130,170	3.83	3.62
75-79 years	378,424	141,317	4.17	3.93
80-84 years	383,869	143,475	4.23	3.99
85-89 years	312,178	118,663	3.44	3.3
90-94 years	176,054	68,321	1.94	1.9
95-99 years	40,837	16,181	0.45	0.45
100+ years	6,352	2,517	0.07	0.07
Ethnic group				
Unknown	1,247,803	509,893	13.75	14.18
White	6,565,710	2,591,536	72.35	72.07
Mixed	118,882	47,465	1.31	1.32
Asian	550,848	220,426	6.07	6.13
Black	336,680	125,136	3.71	3.48
Other	255,005	101,403	2.81	2.82
Quintile of area deprivation				
1 (most deprived)	2,570,020	988,502	28.32	27.49
2	2,038,229	795,764	22.46	22.13
3	1,672,509	665,953	18.43	18.52
4	1,477,398	600,149	16.28	16.69
5 (least deprived)	1,316,772	545,492	14.51	15.17
Arrival mode				
By ambulance	2,727,016	1,120,470	30.05	31.16
Other	6,308,890	2,463,524	69.52	68.51
Not known	38,115	11,866	0.42	0.33
Incident location				
Home	4,523,852	1,886,029	49.85	52.45
Work	343,940	64,366	3.79	1.79
Educational establishment	217,798	12,226	2.4	0.34
Public place	829,448	411,366	9.14	11.44
Other	2,333,164	897,527	25.71	24.96
Not known	826,726	324,706	9.11	9.03

Table 29 continued

Variable	Weekday	Weekend	Weekday	Weekend
	n	n	%	%
Referral source				
General medical practitioner	664,285	84,143	7.32	2.34
Self-referral	5,447,679	2,302,429	60.03	64.03
Local authority social services	7,260	3,236	0.08	0.09
Emergency services	1,185,186	503,780	13.06	14.01
Work	63,524	11,507	0.7	0.32
Educational establishment	36,300	1,798	0.4	0.05
Police	43,560	23,733	0.48	0.66
Other health care provider	345,755	144,913	3.81	4.03
Other	1,183,371	485,441	13.04	13.5
Not known	98,917	34,880	1.09	0.97
Diagnosis category				
Laceration	302,195	138,441	3.33	3.85
Contusion/abrasion	251,376	93,852	2.77	2.61
Soft tissue inflammation	299,473	110,033	3.3	3.06
Head injury	190,573	83,064	2.1	2.31
Dislocation/fracture/joint injury/amputation	444,671	176,916	4.9	4.92
Sprain/ligament injury	352,107	124,776	3.88	3.47
Muscle/tendon injury	128,864	46,027	1.42	1.28
Nerve injury	65,339	25,890	0.72	0.72
Vascular injury	5,445	1,798	0.06	0.05
Burns and scalds	42,652	18,698	0.47	0.52
Electric shock	5,445	2,158	0.06	0.06
Foreign body	68,062	23,733	0.75	0.66
Bites/stings	29,947	15,462	0.33	0.43
Poisoning (inc overdose)	97,102	45,667	1.07	1.27
Near drowning	907	360	0.01	0.01
Visceral injury	2,722	1,079	0.03	0.03
Infectious disease	80,767	36,678	0.89	1.02
Local infection	118,882	51,780	1.31	1.44
Septicaemia	20,872	8,630	0.23	0.24
Cardiac conditions	282,230	95,650	3.11	2.66
Cerebro-vascular conditions	63,524	23,014	0.7	0.64
Other vascular conditions	43,560	13,664	0.48	0.38
Haematological conditions	24,502	7,551	0.27	0.21
Central nervous system conditions (exc stroke)	156,089	56,455	1.72	1.57
Respiratory conditions	297,658	125,136	3.28	3.48
Gastrointestinal conditions	417,447	162,892	4.6	4.53
Urological conditions (inc cystitis)	178,776	72,996	1.97	2.03
Obstetric conditions	27,225	10,788	0.3	0.3
Gynaecological conditions	71,692	28,048	0.79	0.78
Diabetes and other endocrinological conditions	32,670	12,226	0.36	0.34
Dermatological conditions	39,930	19,418	0.44	0.54
Allergy (inc anaphylaxis)	34,485	16,541	0.38	0.46
Facio-maxillary conditions	25,410	13,664	0.28	0.38
ENT conditions	111,622	50,342	1.23	1.4
Psychiatric conditions	87,119	34,161	0.96	0.95
Ophthalmological conditions	99,824	40,993	1.1	1.14
Social problems (inc chronic alcoholism and homelessness)	22,687	10,068	0.25	0.28
Diagnosis not classifiable	1,349,442	531,468	14.87	14.78
Nothing abnormal detected	206,001	84,143	2.27	2.34
Missing	2,996,541	1,181,959	33.02	32.87

6.3.2 Admissions via A&E departments

Results for the population of emergency patients who are admitted to hospital when they attend A&E are given in Table 30. The proportion of the patient population attending A&E at the weekend admitted to hospital compared to those attending during the week is 2.6 percentage points lower. Consequently, average numbers of admissions via A&E are 7% lower for weekend days than for week days.

The risk adjustment model for the probability of admission had a C-statistic of 0.83. The adjusted admission rate of patients attending A&E at the weekend remains significantly lower compared to those attending during the week (odds ratio (OR): 0.946).

The risk adjustment model for the probability of mortality amongst the subset of patients who are admitted when attending A&E had a C-statistic of 0.91. Patients admitted at the weekend have a significantly higher probability of mortality compared to those admitted during the week (OR: 1.054; CI: 1.040 to 1.069). These results are similar regardless of whether risk-adjustment variables are taken from the A&E or inpatient records. Examining the results for each day separately, admissions on Sundays, Saturdays, and Mondays are associated with higher mortality compared to Wednesday admissions. These are the days on which the patients who attend A&E have the lowest risk-adjusted probabilities of admission.

Table 30: Admissions via accident and emergency (A&E) departments and mortality within 30 days

	Average volume of admissions via A&E per day on this day of the week	Crude admission rate on this day of the week	Risk-adjusted admission rate on this day of the week	Crude mortality rate within 30 days following admission via A&E on this day of the week	Risk-adjusted mortality rate within 30 days following admission via A&E on this day of the week (A&E records)[a]	Risk-adjusted mortality rate within 30 days following admission via A&E on this day of the week (APC records)[b]
			Odds-ratio		Odds-ratio	Odds-ratio
Monday	11644.8	28.1%	0.979 [0.974, 0.984]	3.46%	1.032 [1.011, 1.053]	1.036 [1.012, 1.060]
Tuesday	11401.0	30.4%	0.990 [0.985, 0.996]	3.40%	0.997 [0.977, 1.018]	1.000 [0.977, 1.023]
Wednesday	11153.2	30.2%	Reference	3.37%		Reference
Thursday	11241.3	30.5%	1.009 [1.004, 1.015]	3.43%	1.008 [0.987, 1.029]	1.019 [0.995, 1.042]
Friday	11357.5	31.2%	1.010 [1.005, 1.016]	3.43%	0.981 [0.961, 1.001]	1.009 [0.986, 1.033]
Saturday	10557.7	28.4%	0.945 [0.940, 0.951]	3.55%	1.037 [1.016, 1.059]	1.047 [1.023, 1.072]
Sunday	10494.2	26.7%	0.943 [0.937, 0.948]	3.63%	1.081 [1.059, 1.104]	1.088 [1.063, 1.114]
Weekday	11359.6	30.0%	Reference	3.42%	Reference	Reference
Weekend	10525.9	27.5%	0.946 [0.943, 0.950]	3.59%	1.055 [1.042, 1.068]	1.054 [1.040, 1.069]
Difference (Weekend – Weekday)	-833.6 [-940.6, -726.7]	-2.6% [-3.0%, -2.1%]		0.17% [0.08%, 0.27%]		
Ratio (Weekend: Weekday)	0.93	0.92		1.05		

[a] Logistic regression models including controls for age, gender, ethnicity, diagnosis, first or follow-up visit, incident location, accident type, deprivation quintile, month and hospital attended.

[b] Logistic regression models including controls for age, gender, ethnicity, primary diagnosis (SHMI-grouped Clinical Classifications Software category), Elixhauser conditions, admission method, admission source, deprivation quintile, month and admitting hospital. [ ] 95% confidence intervals

6.3.3 Direct admissions

The average number of direct admissions to hospital from services in the community is fairly stable across weekdays, but is 61% lower at weekends (Table 31). The characteristics of patients directly admitted to hospital on weekdays and weekends are given in Table 32. A higher proportion of patients directly admitted at the weekend are children, younger adults, or very elderly (0-34 years or 90 and over) compared to weekdays. The most common primary diagnoses amongst patients directly admitted during the week are abdominal pain, influenza, and headaches. For those directly admitted during the weekend, these are influenza, abdominal pain, and intestinal infections. The population directly admitted at the weekend is less likely to have most of the Elixhauser comorbidities reported.

The average number of patients directly admitted on weekend days and dying within 30 days is significantly lower than for week days (36 versus 81) (Table 31). However, due to the proportionally larger reduction in the average number of direct admissions at the weekend, the proportion of admissions that lead to death within 30 days is higher at weekends than weekdays (2.72% versus 2.37%).

The model used to predict the probability of mortality produced a C-statistic of 0.92. Adjusted mortality rates for directly admitted patients are lowest for Friday admissions (OR: 0.968) and highest for those admitted on Sunday (OR: 1.278). Compared with direct admissions on a weekday, the relative risk of mortality within 30 days was 21.2% higher for direct admissions at the weekend. This equates to a 0.488 percentage point increase in the risk of death, from a baseline of 2.37% during the week.

Table 31: Direct emergency admissions and mortality within 30 days

	Average volume of admissions per day on this day of the week	**Average number of deaths within 30 days of admission per day on this day of the week**	**Crude mortality rate within 30 days following admission on this day of the week**	**Risk-adjusted mortality within 30 days following admission on this day of the week**[a]
				Odds-ratio
Monday	3489.2	83.7	2.40%	1.032 [0.982, 1.085]
Tuesday	3351.4	79.7	2.38%	1.018 [0.968, 1.071]
Wednesday	3232.9	76.5	2.37%	Reference
Thursday	3336.1	78.2	2.34%	0.984 [0.935, 1.035]
Friday	3611.7	85.8	2.38%	0.968 [0.922, 1.018]
Saturday	1397.5	36.7	2.63%	1.154 [1.082, 1.231]
Sunday	1237.3	35.0	2.83%	1.278 [1.196, 1.366]
Weekday	3404.3	80.8	2.37%	Reference
Weekend	1317.4	35.9	2.72%	1.212 [1.162, 1.263]
Difference (Weekend - Weekday)	-2086.9 [-2174.4, -1999.4]	-44.9 [-47.8, -42.0]	0.35% [0.24%, 0.46%]	
Ratio (Weekend: Weekday)	0.39	0.44	1.15	

[a] Logistic regression models including controls for age, gender, ethnicity, primary diagnosis (SHMI-grouped Clinical Classifications Software category), Elixhauser conditions, admission method, admission source, deprivation quintile, month and admitting hospital.

[ ] 95% confidence intervals

Table 32: Descriptive statistics for direct admissions

Variable	**Weekday** n	**Weekend** n	**Weekday** %	**Weekend** %
Observations	817,024	123,835	100.00	100.00
Female	441,520	67,180	54.04	54.25
Age category				
<1 year	51,799	10,699	6.34	8.64
1-4 years	54,414	12,161	6.66	9.82
5-9 years	22,141	4,520	2.71	3.65
10-14 years	19,037	2,923	2.33	2.36
15-19 years	24,184	4,235	2.96	3.42
20-24 years	34,887	6,217	4.27	5.02
25-29 years	36,276	6,056	4.44	4.89
30-34 years	35,704	5,560	4.37	4.49
35-39 years	30,965	4,594	3.79	3.71
40-44 years	34,805	4,966	4.26	4.01
45-49 years	38,809	5,201	4.75	4.2
50-54 years	39,871	5,152	4.88	4.16
55-59 years	39,544	4,842	4.84	3.91
60-64 years	45,672	5,535	5.59	4.47
65-69 years	56,538	6,811	6.92	5.5
70-74 years	55,639	6,737	6.81	5.44
75-79 years	61,195	7,517	7.49	6.07
80-84 years	60,215	8,148	7.37	6.58
85-89 years	46,407	6,972	5.68	5.63
90-94 years	23,612	4,000	2.89	3.23
95-99 years	4,739	879	0.58	0.71
100+ years	572	124	0.07	0.1
Ethnic group				
White	698,147	105,074	85.45	84.85
Mixed	7,435	1,313	0.91	1.06
Asian	35,622	6,068	4.36	4.9
Black	11,030	1,474	1.35	1.19
Other	9,069	1,375	1.11	1.11
Unknown	55,721	8,532	6.82	6.89
Quintile of area deprivation				
1 (most deprived)	191,020	29,671	23.38	23.96
2	165,284	25,176	20.23	20.33
3	165,284	25,052	20.23	20.23
4	155,725	23,095	19.06	18.65
5 (least deprived)	139,711	20,841	17.1	16.83
Admission method				
GP	633,275	102,709	77.51	82.94
Bed bureau	64,627	11,529	7.91	9.31
Consultant clinic	119,204	9,597	14.59	7.75

Table 32 continued

Variable	Weekday n	Weekend n	Weekday %	Weekend %
Bed bureau	64,627	11,529	7.91	9.31
Consultant clinic	119,204	9,597	14.59	7.75
SHMI diagnosis group (138 in total, 10 most common on weekday and/or weekend listed)				
Abdominal pain	48,531	8,235	5.94	6.65
Influenza and other upper respiratory infections…	41,913	8,334	5.13	6.73
Headache and eye disorders	29,658	3,740	3.63	3.02
Complications of fertility and pregnancy	29,331	4,359	3.59	3.52
Skin and subcutaneous infections	26,390	4,161	3.23	3.36
Acute bronchitis	25,164	4,891	3.08	3.95
Other connective tissue disease	24,919	2,340	3.05	1.89
Nonspecific chest pain	24,592	2,142	3.01	1.73
Urinary tract infections	23,040	5,077	2.82	4.1
Hepatitis and sexually transmitted infections	22,386	4,978	2.74	4.02
Intestinal infections	22,223	5,275	2.72	4.26
Pneumonia (excluding tuberculosis and sexually transmitted disease)	22,141	4,012	2.71	3.24
Elixhauser conditions				
Congestive heart failure	30,965	3,926	3.79	3.17
Cardiac arrhythmias	78,598	10,514	9.62	8.49
Valvular disease	21,569	2,365	2.64	1.91
Pulmonary circulation disorders	5,392	582	0.66	0.47
Peripheral vascular disorders	16,831	1,895	2.06	1.53
Hypertension, uncomplicated	179,827	22,340	22.01	18.04
Hypertension, complicated	980	124	0.12	0.1
Paralysis	4,984	842	0.61	0.68
Other neurological disorders	26,717	4,347	3.27	3.51
Chronic pulmonary disease	110,053	14,390	13.47	11.62
Diabetes, uncomplicated	85,216	11,294	10.43	9.12
Diabetes, complicated	9,804	892	1.2	0.72
Hypothyroidism	32,599	4,409	3.99	3.56
Renal failure	62,257	6,353	7.62	5.13
Liver disease	14,788	1,709	1.81	1.38
Peptic ulcer disease excluding bleeding	1,797	248	0.22	0.2
Lymphoma	5,066	854	0.62	0.69
Metastatic cancer	27,125	3,913	3.32	3.16
Solid tumour without metastasis	35,541	5,374	4.35	4.34
Rheumatoid arthritis/collagen vascular diseases	21,079	2,650	2.58	2.14
Congulopathy	4,902	681	0.6	0.55
Obesity	14,461	1,833	1.77	1.48
Weight loss	10,785	1,090	1.32	0.88
Fluid and electrolyte disorders	34,233	5,610	4.19	4.53
Blood loss anaemia	409	50	0.05	0.04
Deficiency anaemia	12,827	1,560	1.57	1.26
Alcohol abuse	19,037	2,402	2.33	1.94

6.4 Discussion

6.4.1 Main findings

Patients admitted to hospital as emergencies at the weekend are known to have a higher rate of death than patients admitted during the week. However, we did not find higher mortality for the whole population attending A&E departments at weekends. The weekend effect was only apparent in the subset of patients who are admitted to hospital, and was far stronger for patients directly admitted from the community – who were admitted in far smaller numbers at weekends – than for patients admitted via A&E. These findings suggest a sicker population of patients is admitted to hospital at weekends, and that this selection effect is partly responsible for the weekend effect.

Elevated mortality rates amongst the population of patients admitted to hospital in an emergency at weekends are driven by a reduction in the volumes of patients admitted to hospital at the weekend rather than an increase in the number of deaths. There were 7% fewer admissions through A&E at weekends, which was not explained by the patient characteristics that we could control for. Hospital staff appear to apply a more stringent admission threshold at weekends to patients seeking emergency care in A&E. This raises the possibility that the patient population admitted at weekends is on average sicker than the population admitted on weekdays, and that this difference is not completely captured by standard risk adjustment using administrative data.

The weekend effect is greatest amongst the patients directly admitted to hospital, for whom the relative risk of mortality was 21% higher at the weekend. However, the number of admissions through this route was 61% lower at weekends compared to weekdays and these admissions represent just 11% (1,317/(1,317+10,526)) of total emergency admissions on a typical weekend day. The lower volume of direct admissions at weekends is not matched by higher A&E attendances or admissions, indicating that patients are not simply being switched between the two routes into hospital at weekends. The concentration of the weekend effect where we see a substantial restriction in the patient flows again raises the possibility that it is due to inadequate measurement of how sick they are rather than lower quality of care at admission.

There may be concern that patients who are directly admitted could experience different quality of care on arrival at hospital. A small proportion (6%) of patients attending A&E are known to have been referred there by a GP, and they are therefore part of the same patient pool as direct admissions in that they initially sought GP care. However, upon arrival at A&E these patients would be expected to receive the same care as those who self-refer to A&E. In an attempt to shed further light on our findings we performed some supplementary analysis on this group of patients. We found that the flows of patients referred to A&E by a GP behaved in much the same way as the direct admissions. The volume of A&E attenders referred by a GP dropped by 68% at weekends, as did the volume of admissions through A&E for this patient group. Attending A&E on a weekend following GP referral was associated with a significantly higher risk-adjusted probability of mortality (OR: 1.168; CI: 1.096 to 1.245). These findings suggest that direct admission to hospital at the weekend is not a cause of elevated mortality, but instead an indicator of an inherently different patient group. If the cause of elevated weekend mortality amongst direct admissions was lower quality of care upon admission rather than referral of sicker patients, we would not expect to see a weekend effect amongst patients referred to A&E by a GP.

6.4.2 Strengths and limitations

This study used data covering the complete population of patients attending consultant-led A&E departments and all emergency admissions to non-specialist acute hospitals in England over an 11-month period. Our risk adjustment models for mortality had high explanatory power (C-statistics equal to 0.92), but in common with previous studies we could not take severity of the primary diagnosis into account, thus limiting our ability to risk-adjust (Lilford et al., 2004). We also did not have information on out-of-hospital deaths and therefore could only include deaths that occurred in any hospital within 30 days of admission. This would generate bias if the proportion of all deaths that occur in a hospital is different for weekend admission.

For an earlier year (1st April 2010 to 31st March 2011), we do have data on out-of-hospital deaths and these show that a slightly higher proportion of all deaths within 30 days of admission occurred in a hospital for weekend admissions (40,614 in-hospital deaths/49,981 total deaths=81.3%) than for weekday admissions (117,989/147,266=80.1%). Our ability to only include in-hospital deaths in the more recent data is therefore likely to have a larger effect in reducing the weekday death rate than the reduction in the weekend death rate. If the weekend and weekday death rates were the same, we would find weekend death rates that were 1.5% higher (81.3%/80.1%) using only in-hospital deaths. Our analysis is therefore likely to contain a small bias towards finding higher death rates at the weekend.

6.4.3 Comparison with previous studies

Previous studies have compared rates of mortality, adjusted for patient characteristics, between those admitted to hospital during the week and their counterparts admitted on weekends (Aylin et al., 2010; Bell and Redelmeier, 2001; Freemantle et al., 2015, 2012; NHS England, Seven Days a Week Forum, 2013b). These studies have consistently found higher mortality rates for patients admitted at weekends, both before and after risk adjustment. Whilst we have also found higher mortality rates amongst patients admitted at weekends, our study differs in two important respects. First, we widened our focus to include all patients attending A&E departments, including those not admitted, in order to avoid possible selection effects in the admitted population. Second, we assessed direct admissions and admissions via A&E separately, in order to gain a better understanding of variations in patient flows throughout the week. Using this approach we found there were fewer patients admitted to hospital in an emergency on weekends, attributable to a 61% lower volume of direct admissions and a 5% relative reduction in the risk-adjusted probability of admission following an A&E attendance. These increased thresholds for admission at weekends are likely to have biased previous studies on weekend mortality.

6.4.4 Policy implications

Current initiatives to move towards seven-day hospital services are only likely to be successful in reducing mortality if reduced availability of services in hospitals on the day of admission is the major cause of the weekend effect. Our findings cast significant doubt over whether this is the case. Patients who attend A&E on weekends are at no higher mortality risk than patients who attend A&E on weekdays. However, a smaller proportion of attending patients are admitted at the weekend and this higher threshold for admission is likely to mean that patients who are admitted via A&E at the weekend are, on average, sicker than patients admitted during the week. Reduced availability of primary care services at weekends means that fewer patients are admitted to hospital via this route and these patients are also likely to be sicker than their counterparts admitted during the week.

Our results add to the increasing body of evidence questioning the use of standardised mortality rates as an indicator of the quality of care in hospitals (Doran et al., 2015; Hogan et al., 2015; Mohammed et al., 2009). The weekend effect identified in previous studies may be a statistical artefact driven by the selection bias introduced by restricting the focus to the admitted population. Extending services in hospitals and in the community at weekends may increase the number of emergency admissions, particularly for patients with less severe illness, and this could have the desired effect of achieving lower hospital mortality rates. However, this would be a statistical phenomenon rather than a clinically meaningful improvement as it would be achieved by admitting less severe patients rather than by reducing the absolute number of deaths.

7. Methods for the economic evaluation of changes to the organisation and delivery of health services: Principal challenges and recommendations

Abstract

Established methods exist for the economic evaluation of new health technologies seeking National Health Service funding in England. Such treatments are subject to rigorous evaluation of cost-effectiveness using standard methodology set out in the National Institute for Health and Care Excellence reference case. This paper is concerned with changes to the organisation and delivery of health services, including changes to health policy, which are not covered by this appraisal process. These changes also have consequences for National Health Service funds, yet undergo no such mandatory cost-effectiveness assessment.

Based on experience of evaluating a regional pay-for-performance programme and a national initiative to extend emergency hospital services at weekends, in addition to reviewing two established frameworks outlining the principles of cost-effectiveness analysis, this paper discusses the principal challenges faced when performing economic evaluations of changes to the organisation and delivery of health services.

The six principal challenges are identified as: undertaking ex-ante evaluation; modelling the counterfactual and estimating the treatment effect; evaluating the impact in terms of quality-adjusted life years; assessing costs and opportunity costs; accounting for spillover effects; and generalisability.

Of these challenges, the methods are currently most advanced in the area of modelling the counterfactual and estimating the treatment effect. There are also methods available for performing ex-ante evaluation, assessing opportunity costs, and examining generalisability. However, these are rarely applied in practice. Methods for estimating the impact on costs and quality-adjusted life years are those most in need of development.

The general principles of assessing the cost-effectiveness of interventions should be applied to all National Health Service spending, not just those involving health technologies. Advancements in the economic evaluation of changes to health services organisation and delivery have the potential to improve the allocation of scarce National Health Service resources.

7.1 Introduction

Established methods exist for the economic evaluation of new health technologies seeking National Health Service (NHS) funding in England. Such treatments are subject to rigorous evaluation of cost-effectiveness using standard methodology set out in the National Institute for Health and Care Excellence (NICE) reference case (NICE, 2013). This thesis is concerned with changes to the organisation and delivery of health services, including changes to health policy, which are not covered by the NICE appraisal process. These will be referred to as 'service interventions' for ease of reading. Service interventions are funded from the same NHS budget as health technologies, yet undergo no such mandatory cost-effectiveness assessment. Whilst attention has grown in recent years around the need for rigorous evaluation of the impact of service interventions, and useful guidelines have been produced (Craig et al., 2008; Lamont et al., 2016; Medical Research Council, 2009, 2008; Raine et al., 2016), these do not extend to economic evaluation. This has resulted in a lack of evidence on the cost-effectiveness of large-scale changes to the organisation and delivery of health services.

The closest thing to a formal evaluation in the realm of health policy is often an impact assessment reported by the Department of Health. Whilst at one point mandated for all new legislation and policy implementation (Shah et al., 2012), it appears that they are no longer compulsory as no such assessment was performed for the recent controversial seven-day services policy (Appleby, 2016). Even when performed, these impact assessments rarely evaluate cost-effectiveness as defined by NICE (Shah et al., 2012). For example, benefits are seldom quantified in terms of quality-adjusted life years (QALYs). The differing levels of scrutiny applied to spending on service interventions as opposed to health technologies is likely to result in allocative inefficiency in the health system.

It is widely accepted that public health is an area which raises additional methodological challenges for economic evaluation, with interventions often targeted at population level (Chalkidou et al., 2008; Weatherly et al., 2009). These include difficulties in: the attribution of effects to an intervention; measuring and valuing outcomes; identifying inter-sectoral costs and consequences; and incorporating equity considerations (Weatherly et al., 2009). Service interventions sit somewhere in between public health and standard health technology assessment (HTA), with changes frequently aimed at an organisational rather than a patient level (Barratt et al., 2016; Ukoumunne et al., 1999). Evaluating these larger-scale changes poses additional methodological challenges above those associated with a typical HTA.

The aim of this thesis has been to contribute to the development of methods for the economic evaluation of changes to the organisation and delivery of health services, using two service interventions as illustrative examples. In this chapter I first briefly outline the NICE HTA process, and the common features of service interventions which make economic evaluation more difficult in this area. I then discuss the six principal challenges faced when performing economic evaluations of service interventions. The specific challenges faced in this thesis were reviewed alongside two established frameworks outlining the principles of cost-effectiveness analysis (Drummond et al., 2015; NICE, 2013) to select the most prominent methodological issues likely to be encountered beyond those generally experienced in an archetypal HTA. The principal challenges are discussed, and recommendations for overcoming them are provided where available.

7.2 Health technology assessment

The NICE technology appraisal process provides recommendations “on the use of new and existing medicines, products and treatments in the NHS” (NICE, 2014). For the purpose of the programme health technologies are defined as: medicinal products, medical devices, diagnostic techniques, surgical procedures or other therapeutic techniques, therapeutic technologies other than medical products, systems of care, and screening tools. Clinical commissioning groups, NHS England, and local health authorities are required by law to comply with NICE technology appraisal guidance that recommends a health technology is made available (NICE, 2014). This rigorous appraisal process ensures that there is a balance between the costs and benefits of healthcare technologies provided by the NHS, ensuring that those funded represent value for money.

As the broad definition of health technologies applied by NICE indicates, there is heterogeneity even within the HTA process. Whilst appearing generic, most international guidelines for economic evaluation were originally designed to assess pharmaceuticals (Drummond et al., 2009; Tunis et al., 2003). As advancements in health technology development have been made, commentators have highlighted additional methodological challenges posed by certain types of technologies which have been overlooked by existing international guidelines. For example, Drummond and colleagues set out six reasons why devices are different from drugs due to a number of inherent characteristics which pose additional methodological challenges for the assessment of both clinical and cost-effectiveness (Drummond et al., 2009). These include the dynamic nature of pricing and the frequent product modifications conducted (such as improvements in software systems and extensions in battery life). Taken together, this means that there is unlikely to be a ‘steady state’ period during which devices could be evaluated in a randomised controlled trial (RCT). Instead, both costs and effects vary substantially over time. The authors argue that whilst the general principles of economic evaluation are still applicable to devices, these additional complexities must be considered, and there is a need for further methodological developments to overcome them.

In the same vein, I have demonstrated in this thesis that whilst service interventions are also inherently different to pharmaceuticals in some important respects, they are still amenable to economic evaluation in some form. These additional complexities may necessitate variations in the methodological approaches taken, but should not be viewed as insurmountable obstacles to the production of robust evidence on the cost-effectiveness of service interventions.

7.3 What makes service interventions different?

As is the case within health technologies, there is vast heterogeneity within service interventions. Despite this variation, service interventions possess a number of common characteristics which make economic evaluation more challenging than for that of pharmaceuticals. Many of these complexities are interlinked, and some present similar issues to those characterising medical devices. Whilst some of these complexities may also be present in certain pharmaceutical products, they are much more common in the case of service interventions, and likely to play a more substantive role in the economic evaluation of service interventions than in an archetypal HTA.

Service interventions are most commonly implemented at an organisational or system level, rather than targeted at an individual (Barratt et al., 2016; Ukoumunne et al., 1999). This can introduce a number of complexities. The causal chain between intervention and effect may be more diluted (Lilford et al., 2010). The effectiveness of a service intervention will be heavily dependent on the interplay of other areas of the system. For example, the cost-effectiveness of a community-based alternative to hospital services will be dependent on the number of referrals that service receives from general practitioners (GPs) elsewhere in the system.

The impact of a service intervention will likely be very context-dependent (Turner et al., 2016), and largely determined by the human component which will differ between organisations. This is in contrast to the mode of action of a drug as an embedded technology, meaning that providing the correct dose is administered, efficacy is dependent on the drug itself rather than the administrator of the drug (Drummond et al., 2009). The complex interactions between different system aspects also means that service interventions are more likely to have both direct and indirect impacts on costs and effects spanning multiple areas of the system, which it is vital to capture to assess the full effects of the intervention in question.

Similarly to medical devices, the effectiveness of a service intervention is also likely to vary over time. The effectiveness may increase as time passes if there is a learning curve involved, or decrease over time if early impacts on outcomes are those easiest to achieve.

Several commentators have challenged the assumptions that RCTs of service interventions are impractical, unethical, and too expensive and difficult to run. Nevertheless, they remain rare (Finkelstein and Taubman, 2015; Haynes et al., 2012). This absence poses a significant practical issue for evaluations in this area. A reliance on observational studies requires more complex analytical approaches when assessing the associated impact on both costs and effects to ensure that the estimates do not suffer from the common issues of selection bias and confounding. The lack of RCTs also means that primary data collection is less common, with a greater need to utilise existing data sources such as those designed primarily for administrative purposes. Whilst adding complexity, the use of observational studies can result in improvements in the external validity of the results (Gillies et al., 2016). The exploitation of secondary data also allows for longer follow-up at a greatly reduced cost compared to RCTs, and so can have advantages (Barratt et al., 2016; Gillies et al., 2016).

7.4 Principal challenges faced when conducting economic evaluations of service interventions

7.4.1 Ex-ante evaluation

Of the evaluations of service interventions which are performed, the majority are ex post. Ex ante modelling of the potential costs and benefits of a proposed service change can, however, be very informative and should be utilised to a much greater extent than is currently the case. In the same way that a trial and/or simulation model would be undertaken before introducing a new treatment to the NHS, ex ante modelling can be used to estimate the likely impact of the planned service change based on the evidence available. This was the purpose of the impact assessments previously performed by the Department of Health. Such ex ante analysis should be used to select from potential service changes to be pursued, revealing those which could not possibly be cost-effective even under the most optimistic assumptions. This would save substantial NHS resources from being wasted implementing changes which, even if completely successful in achieving their aims, could never represent a cost-effective use of scarce resources. Ex ante modelling could also be used when designing potential service changes, revealing targeted areas with the potential to offer cost-effective solutions.

Ex ante modelling can range from a relatively simplistic exercise estimating the maximum potential benefits of the proposed change in terms of QALYs to a detailed simulation model, and can serve a number of different purposes. Such detailed microsimulation models are commonly employed by the Congressional Budget Office in the United States of America (USA) to examine the potential impacts of proposals to increase health insurance coverage (Congressional Budget Office, 2007).

This ex ante modelling should form the initial stage of any planned service change, and could often highlight just how sparse the evidence base for many planned programmes is. If the service intervention in question is targeted at patients with a particular condition, then estimates of the current burden of disease could be utilised to obtain estimates of the potential health gain. Evidence on the incidence of the condition or event in question could give potential effect size estimates, and be used to calculate the possible resource use implications.

If the budget set to be allocated to the programme is known, ex ante modelling could be used to estimate the threshold of improvement that would need to be achieved for it to be cost-effective. Any available literature on the expected effect size could then be compared to this to assess whether the service intervention is likely to produce benefits of that magnitude. Finally, ex ante modelling could be used to set the maximum amount of money which should be allocated to proposed service changes, so that the cost of a programme should not exceed a pre-defined budget if it is to remain a cost-effective use of resources.

The ex-ante analysis of the proposed plans to increase weekend hospital services performed in Chapter 5 provides an example. The service change was proposed in response to the finding that the risk of mortality is higher amongst patients admitted to hospital at the weekend compared to their counterparts admitted during the week. Yet closer examination of the evidence revealed that even if the mortality rate experienced by patients admitted to hospital in an emergency at weekends was reduced to that of patients admitted during the week, the costs of the planned service extensions would exceed the maximum that NICE would recommend the NHS should be prepared to pay to achieve a health gain of this size. Yao and colleagues provide a framework for the ex-ante evaluation of generic service delivery interventions, and road-test this on an intervention to improve patient handover of care between hospital and the community (Yao et al., 2012). Their nine step guide to prospective evaluation includes the identification of suitable endpoints, the use of expert elicitation in the absence of evidence on the likely effectiveness of the intervention, methods for deriving and synthesising data to input into an economic model, and calculations of cost-effectiveness. Notably, they suggest presenting the cost-effectiveness estimates in a number of different ways to aid decision makers and reflect the uncertainty surrounding the estimates.

The framework suggests using the ‘headroom’ method, which explores the minimum health benefits that an intervention must produce for it to be considered cost-effective at a given societal willingness-to-pay value. This magnitude can then be compared to the point estimate and confidence intervals of the estimated effect size to assess the likelihood that the intervention will be cost-effective. The authors also present the costs and benefits associated with the intervention if it were to be 100% effective in achieving its aim of reducing preventable adverse events, and the effectiveness level at which the intervention becomes cost-saving or ‘dominant’. Providing the estimates for these various scenarios may be more informative to decision makers than a single point estimate. The authors state that this is the first study known to them which attempts an ex ante economic evaluation of a proposed service delivery intervention, highlighting the lack of economics research conducted at the design and development stage of service interventions.

7.4.2 Modelling the counterfactual and estimating the treatment effect

A clear comparator is essential for any economic evaluation, representing what would have happened in the absence of the programme (Drummond et al., 2015). Despite calls for the use of RCTs in this area, most service interventions are introduced in a non-experimental fashion (Finkelstein and Taubman, 2015; Haynes et al., 2012). This creates challenges in modelling the counterfactual, and attributing any effects detected to the service change of interest. Unlike a trial, the evaluator often has no involvement with the implementation of the service intervention in question, and so must design the evaluation ex post with no control over the exposure to treatment.

Endogeneity is a particular problem, with selection into treatment frequently correlated with expected outcomes. For example, the Ashenfelter’s dip often observed before treatment which can occur when agents have self-selected into treatment on the basis of the expected associated improvement in outcome (Ashenfelter, 1978). The latest Medical Research Council (MRC) guidance on the use of natural experiments helpfully summarises the suite of methods available to deal with such selection, including instrumental variables, difference-in-differences, regression discontinuity, and matching (Craig et al., 2012; Medical Research Council, 2009).

Observation of changes in outcomes prior to the implementation of a new service intervention can, conversely, indicate anticipatory effects when announcement of a forthcoming scheme induces changes in behaviour before the programme is officially introduced (Malani and Reif, 2015). It is vital for evaluators to determine whether such changes indicate problems of endogeneity or anticipatory responses. Failure to capture anticipatory responses can invalidate the analytical approach taken, and lead to underestimation of the true effect of the programme if some initial impacts are missed. This is particularly important if using difference-in-differences methodology as such effects impact upon the pre-trends, and in turn the estimated treatment effects. Key to specification of the counterfactual is therefore an understanding of how the changes take effect over time. Malani and Reif provide a useful framework for rigorously comparing and estimating the various models that may be used to estimate anticipation effects based on economic theory (Malani and Reif, 2015).

A recent evaluation of a change to the way hospitals were paid for certain daycase surgeries in England identified rapid responses to the new financial incentives shortly following the announcement of the scheme, which was four months before its formal introduction (Allen et al., 2016). Instead of the usual pre and post design, the evaluation was therefore divided into three periods (pre, anticipatory, and post) to formally capture these effects and ensure that the pre-treatment period was not contaminated by the announcement of the payment change.

In addition to understanding any anticipatory effects prior to the formal implementation of a service intervention, it is also necessary to consider how the impacts of the programme may evolve over time. The effects of a service intervention are likely to differ between the short- and long-run, either because changes take time to come into effect (for example if there is a learning curve involved) or because improvements may reach a ceiling after which no further gains can easily be achieved. This is particularly relevant to quality improvement initiatives.

Defining the relevant counterfactual is more complex for a long-term evaluation, and some of these complexities are illustrated in the long-term evaluation of the Advancing Quality (AQ) initiative presented in Chapter 3. The short-term evaluations of the programme found that the introduction of AQ led to a significant reduction in mortality for participating hospitals during the first 18 months of the scheme (Sutton et al., 2012), and that the resulting health gain represented a cost-effective use of NHS resources (Chapter 2).

When returning to conduct the longer-term evaluation over a further 24 months the relevant comparator was not immediately clear. We could have evaluated mortality over the longer-term period against the trend in mortality before the programme's introduction, which would have examined whether mortality was reduced over the longer-term period compared to the counterfactual situation in which AQ had never been implemented. However, given that the programme had been introduced and altered the trend in mortality rates over the first 18 months, we felt that the relevant decision problem was not whether AQ was effective overall but whether it generated additional impacts beyond the short-term period already evaluated. This approach assumed that the initial impact of AQ was a permanent change to the trend in mortality, which would have persisted had the programme been stopped after the first 18 months. Neither approach is definitively correct, but it is important to be aware of the different potential decision problems that can be answered by a long-term evaluation and define the counterfactual accordingly.

The lack of random treatment assignment in the context of service evaluations also necessitates a greater deal of sophisticated risk-adjustment be performed as part of the analysis, to ensure that any treatment effects detected are not due to other confounding factors. The re-examination of the impact of weekend emergency hospital admission in Chapter 6 illustrates the dangers of inadequate risk-adjustment. Previous interpretations of studies detecting elevated mortality amongst patients admitted to hospital at weekends had concluded that this effect was the result of poor quality care. Yet the subsequent analysis presented in Chapter 6 suggested that these results reflect the inability of previous studies to fully control for patient severity, with the population of patients admitted to hospital at weekends sicker than those admitted during the week. This more robust analysis suggested that the previously detected weekend effect may be merely a statistical artefact driven by selection effects and insufficient risk-adjustment, rather than a matter for policy concern.

7.4.3 Evaluating the impact in terms of QALYs

The primary outcomes assessed in service evaluations are often intermediate process, clinical or patient safety measures. Readmissions, length of stay, and mortality rates are also commonly evaluated. Whilst these are important indicators of a programme's success, the value of the effects of service interventions should also be measured in terms of their impact on health outcomes. Economic evaluation typically requires these to be expressed in terms of QALYs, as is favoured by governmental agencies in a number of countries including the United Kingdom (UK), Canada, and Australia, (Australian Government Department of Health, 2015; Canadian Agency for Drugs and Technologies in Health, 2006; NICE, 2013). This allows decisions to be made over whether the service intervention should be funded based on comparisons on the same terms as other claims on healthcare resources.

Service evaluations tend to rely on administrative data. Whilst containing a wealth of useful information, administrative data often does not contain the necessary preference-based health-related quality of life measures needed to estimate QALYs. An exception to this is the English patient reported outcome measures (PROMs) programme, which covers unilateral hip and knee replacements, varicose vein surgery, and groin hernia repairs (Department of Health, 2008; Gomes et al., 2016). All providers of NHS-funded inpatient care have been required to collect PROMs from patients undergoing these four elective procedures since April 2009. The data collected includes the EQ-5D, which patients are invited to complete both before surgery and either three or six months after their operation. This allows for the direct calculation of QALYs which could then be used in an evaluation of a service intervention targeting any of these four conditions. It is important to note, however, that the PROMs programme only covers individuals who choose to undergo surgery. Some individuals will opt not to have an operation, but no information is available on the health of those choosing this route for use as part of the counterfactual.

In the absence of such PROMs data, a method to estimate the impact of AQ on QALYs indirectly using administrative data was developed in Chapter 2. This method involved estimating the impact of the programme on mortality, and applying a discounted and quality-adjusted life expectancy (DANQALE) tariff to these mortality changes to estimate the impact of AQ on QALYs for the patients affected. The DANQALE tariff is stratified by single year of age (18 – 100 years) and sex. Sex-specific life expectancy estimates at each single year of age are then taken from the Interim Life Tables produced by the Office for National Statistics (ONS) (Office for National Statistics, 2011a). This provides a method for estimating the effects of a programme in terms of QALYs which could be easily applied in other service intervention evaluations using widely available administrative data. Chapter 4 further developed the methods for estimating the length of life component of the DANQALE method using survival analysis techniques commonly employed in clinical trials.

It is also possible to supplement analysis conducted using administrative data with economic modelling techniques, as demonstrated in an evaluation of the centralisation of stroke services in London (Hunter et al., 2013). Hunter and colleagues utilised audit data containing a condition-specific health measure, and derived utility values for health states based on this information using a mapping algorithm. Following this, they built a Markov model to estimate the incremental cost-effectiveness of the service reorganisation using a combination of information available in Hospital Episode Statistics and three stroke audits.

These examples demonstrate that it is possible to obtain estimates of a programme's impact in terms of QALYs even in the absence of primary data collection, and similar methods should be utilised to a much greater extent in order to evaluate the value for money offered by service interventions. The wealth of published algorithms mapping condition-specific measures to preference-based measures of health-related quality of life could provide a solution to the absence of such utility measures in many cases (Brazier et al., 2010). There is, of course, much room for further improvement in the methods to estimate QALYs in the absence of primary data collection. Such developments would make a valuable contribution to this field.

7.4.4 Assessing costs and opportunity costs

As described by Maynard, although effectiveness measurement is essential, "it is like a cart without a horse if it is not matched up with cost data, which demonstrates how much care is given up when a procedure is adopted" (Maynard, 2012. p.10). HTA is concerned with the mean cost across treatment arms, and the average incremental cost of the treatment compared to the control (NICE, 2013). Whilst the majority of service evaluations conducted to date have focused on effectiveness, those which have estimated costs have generally examined a wider range of questions than those addressed in a typical HTA. These include investigating how costs vary at the margin, how the scale of implementation impacts costs, and quantifying the total cost of the intervention rather than the incremental costs at patient-level. Others have ignored effectiveness entirely, solely focusing on whether services are cost saving.

Costing consists of three elements; identification, measurement, and valuation of the resources associated with each programme under consideration (Drummond et al., 2015). Economic evaluations conducted alongside clinical trials are able to collect primary data on patient-level resource use as part of the study. Identification and measurement of the relevant resources is likely to be more complex when examining service interventions, not least because of the common reliance on secondary data. The complex and interlinked nature of health service organisation means it can be much more difficult to assign resources to particular services, with many resources not being fully divisible. In addition, patient-level resource utilisation may not be available, with evaluators instead forced to rely on the provision of programme-level cost estimates. A top-down costing approach may be the only option, with micro-costing often unfeasible.

When it comes to the valuation stage, evaluations of service interventions can make use of national casemix-related costs for episodes of care such as diagnostic-related groups (DRGs) or healthcare resource groups (HRGs). These can be taken as reasonable approximations for the costs of providing treatment for different categories of patients (Drummond et al., 2015). Such tariffs represent average treatment costs across all patients and providers, with the marginal costs from a change in activity likely to differ from the provider's perspective. However, from the perspective of the commissioner these are the actual prices paid, and so do represent the appropriate cost figures to apply from this evaluation perspective. Average costs are also thought to be more appropriate when considering matters of national policy, as they reflect the true variable costs when many services are delivered by a large number of providers across the country (Drummond and Jefferson, 1996). In the long-run, average costs are taken to equal to marginal costs, as all resources are assumed to be variable (Knapp et al., 1990).

Administrative data can also be utilised to estimate impacts on costs through examining intermediate outcomes such as length of stay and readmissions. This was the approach taken in Chapter 2 when evaluating the cost-effectiveness of AQ. In addition to examining the costs of setting up and running the programme, per diem tariffs were applied to length of stay and emergency readmissions to examine the cost-consequences of any effects on these events. In doing so it was vital to fully understand the payment arrangements in operation to ensure that any detected effects did represent cost changes from the perspective of the evaluation. For example, due to the way in which hospitals are reimbursed for admission episodes in England, cost savings from reducing length of stay are often reaped by providers, with commissioners paying the same tariff regardless of length of stay below the trim point. It is therefore important to consider who any such cost implications fall upon.

Drummond and colleagues conclude that whilst cost analysis is a central feature of economic evaluation, it has received relatively little attention from analysts to date (Drummond et al., 2015). The focus in this area has been on developing appropriate statistical methods to deal with the distributional characteristics of cost data, with a neglect to the costing process itself (Graves et al., 2002). This is all the more true in service evaluation, where costs are often not considered at all.

As with any economic evaluation, the objective of placing a monetary value on the resources utilised is to obtain an estimate of their opportunity costs. In the case of healthcare spending this represents the possible health gains foregone through not providing alternative treatments. Regardless of whether a service change is financed by additional funds or involves a reallocation of current resources, there will still be opportunity costs in terms of potential care displaced. When making the decision over whether a new service intervention should be funded, it is therefore vital to consider how much health will be displaced as a result.

Considerations of opportunity cost are taken into account in economic evaluation using the cost-effectiveness threshold, which is taken to represent an estimate of the health forgone elsewhere as other NHS activities are displaced (Claxton et al., 2015). However, the current threshold of £20,000 to £30,000 per QALY employed by NICE in England was founded on the values implied by past decisions rather than evidence of the true opportunity costs experienced in the system. Recent work conducted by Claxton and colleagues produced the first empirical estimates of the true scale of opportunity costs faced by the NHS in England when additional costs are imposed (Claxton et al., 2015). The authors estimated the cost-effectiveness threshold to be £13,936 per QALY. This implies that funding any new interventions with an incremental cost-effectiveness ratio (ICER) higher than this figure imposes a net loss on the system, with the health lost from displacement larger than the health gained from the intervention in question.

It is unclear what impact this work will have on the threshold used by NICE in the future. Regardless, the research provides a methodology for estimating the true opportunity costs associated with funding decisions. Employment of this methodology in service evaluations would enable a quantification of the magnitude of health forgone and an indication of where this opportunity cost is likely to fall, making the rather abstract notion of opportunity costs somewhat more tangible to decision makers.

7.4.5 Accounting for spillover effects

Service interventions can have consequences beyond their intended effects. A change in one area may lead to an unintended diversion of effort away from other areas of care not covered by the initiative, having a negative impact. Alternatively, changes aimed at improving one area of a service may lead to improvements in care in other areas, having a positive impact. These additional effects are often termed spillovers, defined as impacts on costs and/or outcomes other than those explicitly targeted by the intervention. Spillover effects can occur within patients targeted by the intervention onto other untargeted areas of that patient’s care, across patients not targeted but treated within the same organisation, or across organisations (Table 33). The form of spillover can be classified by the level of agent over which it occurs, and whether the spillover occurs within or between agents. A key distinction is between tasks which are substitutes and complements. In contrast to opportunity costs, which represent the indirect impact on unidentified patients as a result of possible healthcare foregone elsewhere in the system, spillovers are direct impacts which can be identified by widening the evaluation's scope.

Table 33: Classification of spillover effects

Level	**Within**	**Between**
Patients	Targeted patients experience changes in treatments not included in the intervention	Un-targeted patients experience changes in treatment
Professionals	Professionals respond to the service change by changing their performance of other tasks	Professionals not included in the service change respond to the service change
Organisations	Organisations re-prioritise their internal allocation of resources	Organisations respond to the performance of others

These spillover effects introduce two additional complexities into evaluations of service interventions. Firstly, the potential for spillover effects means that it may be necessary to look beyond the intended outcomes of a programme in order to assess the full impact of the service change. Spillovers onto other patient groups or areas of care could impact on the cost-effectiveness decision of an initiative if additional benefit or cost consequences occur beyond the intended scope. Secondly, the possible existence of spillovers requires consideration when selecting an appropriate control group in a non-experimental study design as it is vital that the control group is truly unaffected by the change under examination.

The Quality and Outcomes Framework (QOF) introduced financial incentives for certain areas of care delivered by GPs, providing a good example of where spillover effects may be relevant. GPs received financial incentives for keeping records of risk factors such as smoking status for patients with 10 chronic diseases. One evaluation of the scheme explicitly looked for evidence of spillovers from the incentivised to un-incentivised activities (Sutton et al., 2010). Sutton and colleagues considered two potential forms of spillovers: horizontal spillovers capturing the effect on non-incentivised activities for the patients targeted by the scheme, and vertical spillovers covering the effect for untargeted patients on activities which were incentivised for other targeted groups of patients. As noted in Chapter 2, these effects are likely to differ because the mechanisms through which they might arise are different.

In the case of the QOF, horizontal spillovers can arise through more comprehensive care in single consultations (e.g. interventions that address multiple lifestyles) whilst vertical spillovers may arise if primary care practises introduce new systems for particular activities (e.g. recall systems for recording of smoking status) that are then applied across the wider practice population. Sutton and colleagues found evidence of substantial positive horizontal spillovers, with recording of clinically effective but un-incentivised risk-factors such as alcohol consumption increasing for patients who were targeted by the QOF (Sutton et al., 2010). As the QOF scheme paid GP practices for the recording of risk factors, the inclusion of these positive spillover effects reduced the cost to commissioners per risk factor recorded by a factor of two.

In Chapter 3, evidence of spillovers resulting from the AQ programme were also detected. There was some evidence to suggest that AQ had generated positive spillovers over the longer-term both for patients not targeted by the programme but treated by the same doctors as those treating incentivised patients, and for patients treated for incentivised conditions in other hospitals not directly involved in the original initiative. Whilst the detection of these spillovers was a positive finding in that the programme had larger effects on quality than was originally anticipated, a number of supplementary analyses were required as a result to ensure that the control group was still a valid representation of the counterfactual.

Economic evaluation methods go some way towards taking account of negative spillovers by considering the opportunity costs of resources used by the intervention (Chapter 2). However, current approaches do not consider that negative spillovers could take other forms and do not take positive spillovers into account at all. Leaving these dimensions of spillovers out of economic evaluations could distort conclusions about a service intervention's cost-effectiveness, leading to allocative inefficiency. Further research is required to develop a systematic approach to the identification and measurement of spillovers in service intervention evaluations (Kristensen et al., 2015).

7.4.6 Generalisability

Whilst it may be of some value to know whether a particular service intervention represented a cost-effective use of resources in the local setting within which it was implemented, once the change has been made there may be little ability to reverse it even if it is later shown to not be cost-effective. Furthermore, the effectiveness of a service is likely to be context-dependent, and largely determined by contextual factors which will differ between organisations (Turner et al., 2016). Of greater use to decision makers is generalisable evidence on service changes and the potential impacts of scaling and spreading service interventions which are found to be cost-effective. This poses a different decision problem to the typical question addressed in a HTA, of whether option A is cost-effective compared to option B.

Whether a particular service intervention is cost-effective is the simpler question to answer, as this does not require understanding of the production process, but simply an estimate of the treatment effect and costs associated with the programme (Shiell et al., 2008). However, an understanding of which elements of a programme generate positive treatment effects, and how these could be implemented elsewhere, are far more informative (Lamont et al., 2016). This requires an understanding of the underlying causal mechanism, or 'active ingredients' of a programme as discussed in the MRC guidelines for developing and evaluating complex interventions (Medical Research Council, 2008).

It may not be possible to unpick the causal mechanisms behind an intervention using quantitative analysis alone. A mixed methods approach may therefore be necessary, combining the quantitative measurement of costs and benefits with a qualitative exploration of how and why impacts are (or conversely are not) generated (Turner et al., 2016). This was the approach taken in the AQ evaluation, where qualitative work undertaken as part of the wider project explored the potential explanations for the positive impacts of the programme, which were in contrast to those of many previous pay-for-performance initiatives. This work found that a number of factors appeared to contribute to the success of AQ, including collaborative learning events and dedicated infrastructure support, in addition to the accompanying financial incentives (McDonald et al., 2015). The mixed-methods approach enabled generalisable lessons to be produced for the wider implementation of pay-for-performance schemes across the NHS as a whole, learning from the success of AQ. A recent paper by Anderson and Hardwick proposes a means by which economic evaluations could be combined with theory-driven realist evaluations to better explain how service interventions work, who they work for, in what circumstances, and why (Anderson and Hardwick, 2016).

7.5 Discussion

Changes to the organisation and delivery of health services are not subject to the same assessment of cost-effectiveness mandated for all new healthcare technologies in England. The differing levels of scrutiny applied to spending on service interventions as opposed to health technologies is likely to result in allocative inefficiency in the health system.

Whilst the general principles of economic evaluation are still applicable to service interventions, their inherent characteristics pose additional complexities that demand consideration from evaluators. This chapter aimed to discuss the principal challenges faced when performing economic evaluations of service interventions, and provide recommendations for overcoming them where available. Although the need for rigorous evaluation of the impact of service interventions has been acknowledged, and many important advancements made over recent years (Craig et al., 2008; Lamont et al., 2016; Medical Research Council, 2009, 2008; Raine et al., 2016), this paper represents the first extension of this discussion to the field of economic evaluation.

The discussion presented draws on experience of examining the cost-effectiveness of two prominent service interventions as part of this thesis, in addition to insights offered by the wider literature. Considerations are debated around ex-ante evaluation, modelling the counterfactual and estimating the treatment effect, evaluating impact in terms of QALYs, assessing costs and opportunity costs, accounting for spillover effects, and generalisibility.

For the purpose of illustration, pharmaceutical HTAs and service interventions have been dichotomised. In reality all interventions sit somewhere on a scale, with many of the technologies evaluated through the NICE HTA programme containing service elements. All economic evaluations face challenges to some extent. However, service interventions possess a number of common characteristics which make economic evaluation more challenging than for that of drugs.

Emphasis has been placed on measuring the benefits of service interventions in terms of health in this discussion, as quantified using QALYs. Whilst the QALY is not without criticism, its explicit characterisation of health and societal preferences for such health provides support to decision makers, promoting consistency and transparency across funding decisions (Drummond et al., 2015; Reed Johnson, 2009; Smith et al., 2009). This is not to say that other potential impacts of service interventions are immaterial. Cost-effectiveness should be considered in addition to, not in replacement of, other outcomes. These could include impacts on waiting times, patient satisfaction, and inequalities. Cost-effectiveness is a key, and often dominant, consideration in the NICE appraisal process, but it is not the sole factor determining decision-making (Shah et al., 2012). In the same vein, whilst cost-effectiveness may not be the only factor relevant for service evaluations, it should be a major consideration.

Of the challenges discussed in this paper, methods are currently most advanced in the area of modelling the counterfactual and estimating the treatment effect. Much effort has been devoted to overcoming the issues faced by non-random allocation of treatment, with numerous authoritative guides summarising the suite of methods available and situations in which they should be applied (Craig et al., 2017, 2012; Gillies et al., 2016; Medical Research Council, 2009). There are also methods available for performing ex-ante evaluation, assessing opportunity costs, and examining generalisability. These are, however, rarely applied in practice. Whilst further research would be beneficial in all areas highlighted, methods for estimating the impact of service interventions on costs and QALYs are those most in need of development.

It is clear that the methods of HTA cannot (and should not) simply be applied straight to service evaluations, and that new methodological approaches and innovations are needed. However, the general principals of assessing the cost-effectiveness of interventions should be applied to all NHS spending, not just that which falls on health technologies. Both strands of health economics have made impressive methodological progress in different aspects of evaluation, and could learn valuable lessons from each other. Whilst there is still much progress to be made in the economic evaluation of service interventions, advancement in this area has the potential to greatly increase allocative efficiency in the health system, improving the allocation of scarce NHS resources.

8. Discussion

8.1 Overview

The overarching aim of this thesis was to contribute to the development of methods for the economic evaluation of changes to the organisation and delivery of health services. Two example service interventions were examined, and specific evidence on the costs and outcomes associated with these was provided. This thesis illustrates that whilst service interventions are inherently different to pharmaceuticals in some important respects, they are still amenable to economic evaluation. Service interventions pose additional complexities for economic evaluation, which may necessitate variations in the methodological approaches taken. These additional complexities are reviewed, and methodological advancements developed.

This final chapter will summarise the key findings of this thesis, and discuss the strengths and weaknesses of the research presented, in addition to the implications of this work, and directions for future research in this area. As this thesis is presented in journal format, the specific findings, strengths, and limitations are already discussed in each chapter. This section therefore discusses the overarching themes of the thesis, and the main limitations common across chapters.

8.2 Summary of findings

This thesis aimed to pursue four objectives. The main findings are therefore summarised in relation to these objectives below.

8.2.1 Reviewing and critiquing the existing literature pertaining to the economic evaluation of the two example service interventions studied

Chapter 2 began with a systematic review of previous economic evaluations of pay-for-performance (P4P) programmes. This review found that existing evaluations of P4P programmes had focused on the impact of these schemes on the targeted quality measures, with the majority neglecting the more pertinent issue of their effects on health outcomes and costs. The quality of the 14 studies that did attempt to examine costs was generally found to be poor, with the costs included in many of the studies inconsistent with their stated perspectives. Most of the reviewed studies failed to incorporate the full range of relevant costs, with the range of effects considered also found to be narrow. Conclusions regarding the value for money offered by previous P4P programmes could therefore not be made.

In Chapter 5 the evidence base being used to support the case for seven-day hospital services in England was critiqued. The existing evidence was found not to show sufficient support for the likely effectiveness and cost-effectiveness of the planned service extensions. It was noted that many previous studies estimating the magnitude of the weekend effect failed to report the baseline level of mortality risk, instead only presenting the magnitude of the weekend effect in terms of the relative risk increase. Relative risks alone are uninformative as the absolute magnitude of the effect cannot be determined. Evidence on the cause of the previously detected weekend effect and the ability of service extensions to reduce it was also found to be lacking. This appraisal therefore concluded that there is as yet no clear evidence that seven-day working will reduce the weekend death rate.

Further critical assessment of the approach taken to estimate the weekend effect in previous research was provided in Chapter 6. The possibility that earlier estimates of the weekend effect are subject to bias because they focus only on admitted patients was explored. The empirical investigation of this hypothesis demonstrated this to be the case, suggesting that the weekend effect in mortality amongst the admitted population may reflect admission of fewer and sicker patients who are at greater risk of dying.

8.2.2 Developing methods for the economic evaluation of service interventions

Following the systematic review of previous economic evaluations of P4P programmes, a more comprehensive analytical framework was developed in Chapter 2 to guide the assessment of the cost-effectiveness of P4P schemes. This framework adapted the National Institute for Health and Care Excellence (NICE) reference case, highlighting the issues that should be considered in relation to the perspective of evaluations, the comparator, cost categories to be included, opportunity costs, the outcomes to be evaluated, and the relevant time horizon.

Chapter 2 also developed a new method for quantifying the effects of service interventions in terms of quality-adjusted life years (QALYs) in the absence of primary data on health-related quality of life. The discounted and quality-adjusted life expectancy (DANQALE) tariff proposed provides a methodology by which changes in mortality attributable to a service intervention can be converted to QALY gains, using age-sex specific life expectancy estimates and mean EQ-5D values reported by respondents to the Health Survey for England.

This method for quantifying the effects of service interventions in terms of QALYs was then revisited and further developed in Chapter 4. The focus of Chapter 4 was on developments in the estimation of the length of life component of the QALY, accounting for the fact that the length of life of the patients affected by health care policies and programmes is likely to differ from that of the general population. This work demonstrates how survival analysis techniques commonly employed in clinical trials can be used to improve treatment effect estimates associated with policy initiatives.

Chapter 5 developed methods for ex ante evaluation of service interventions, illustrating how existing evidence can be utilised to estimate the potential costs and benefits of a planned service change before it is implemented. This evaluation again utilised the DANQALE method to estimate the potential impact of seven-day hospital services in terms of QALYs, illustrating how the method can be applied ex ante as well as ex post.

Chapter 6 focused on estimating the treatment effect associated with weekend admission, showing that earlier estimates suffer from bias due to the restricted focus on only the admitted patient population. Previously under-utilised data on accident and emergency (A&E) department attendances was used to widen the focus of the analysis, offering a solution to this issue.

8.2.3 Producing evidence on the costs and benefits associated with two example service interventions

Chapters 2, 3, and 4 provided evidence quantifying the benefits of the Advancing Quality (AQ) P4P programme, whilst Chapter 2 also quantified the scheme's cost.

Chapter 2 presented an initial estimate of the impact of AQ during the first 18 months of the programme. AQ was found to have led to a statistically significant reduction in mortality and length of stay. The mortality impacts translated to an estimated reduction of 649 deaths and a gain of 5,227 QALYs. These benefits were achieved at a total cost of £13m to commissioners. The costs of AQ were therefore far lower than the £105m the National Health Service (NHS) would be willing to pay to achieve a health gain of the size of that demonstrated by the programme, as evaluated at the value of £20,000 per QALY used by NICE.

As part of the methodological developments in Chapter 4, a subgroup of the patients affected by AQ were examined. The impact of AQ on life years gained was estimated amongst patients admitted for pneumonia during the first 12 months of the programme's introduction. AQ was estimated to have been associated with a gain of 0.38 years per patient. This compares to an estimate of 0.154 years per patient when the impact of the programme was assessed using the previous method of applying general population life expectancy figures to those surviving past 30 days after admission. The increase in treatment effect estimate produced using survival analysis suggests that AQ impacted on survival past the 30-day post-admission window usually assessed.

The impact of AQ on mortality over the longer-term was assessed in Chapter 3. Whilst mortality for patients admitted to hospital Trusts in the North West for the conditions covered by AQ continued to fall throughout the 42 months assessed, the reduction in mortality was greater for Trusts in the rest of England than in the AQ region during the longer-term period (months 19 - 42). The short-term relative reductions in mortality detected in Chapter 2 were therefore not maintained over the longer-term.

Chapter 5 estimated the costs and benefits of implementing seven-day hospital services. The maximum achievable health gain possible if the mortality rate amongst patients admitted to hospital in an emergency at weekends was reduced to that of patients admitted during the week was estimated to be between 29,727 and 36,539 QALYs per year. Using the NICE threshold of £20,000 per QALY, the NHS should spend no more than £595m to £731m to achieve a health gain of this size. The cost of implementing the planned seven-day services programme was estimated to be between £1.07bn and £1.43bn. This cost exceeds the maximum amount that the NHS should spend to eradicate the weekend effect, meaning that there is no evidence to suggest that the planned service extensions would represent a cost-effective use of resources.

8.2.4 Identifying the principal challenges faced when conducting economic evaluations of service interventions

Informed by the evaluations undertaken in the proceeding chapters and two established frameworks outlining the principles of cost-effectiveness analysis, Chapter 7 sought to identify the principal challenges faced when conducting economic evaluations of service interventions. Considerations around ex-ante evaluation, modelling the counterfactual and estimating the treatment effect, evaluating the impact in terms of QALYs, assessing costs and opportunity costs, accounting for spillover effects, and generalisibility, were identified as the six most prominent methodological challenges likely to be faced when performing economic evaluations of service interventions. Recommendations for overcoming the key challenges faced were provided where available.

8.3 Strengths and weaknesses

8.3.1 Strengths

The two service interventions examined illustrate differing characteristics, providing contrasting examples upon which to further develop methods for the economic evaluation of changes to the organisation and delivery of health services. The quasi-experimental nature of AQ's introduction, its focus on a small number of specific patient groups, and the ex-post approach taken made this programme a useful foundation upon which to begin the methodological developments. The national coverage of the seven-day services policy across all patients admitted to hospital in an emergency at weekends, and the ex-ante approach taken increased the complexity of this evaluation, tackling more complex challenges. The ex post examination of AQ enabled the effects of the programme to be estimated directly, whilst the ex-ante evaluation of seven-day services required prediction of the policy's costs and benefits based on the evidence available.

The example service interventions studied represent high-profile programmes in England, covering large populations, and involving substantial resources. In addition to the contribution to the development of methods in this thesis, the evaluations of these two schemes provide important evidence on the likely cost-effectiveness of these large-scale initiatives. The evaluations presented in this thesis are the first to provide estimates of the costs of these programmes, and of the benefits associated with these two service interventions quantified in terms of QALYs.

All five empirical chapters of this thesis utilised inpatient records from Hospital Episode Statistics, a rich patient-level data set covering the entire population of patients admitted to hospital in England. This was supplemented with data linked to deaths occurring out of hospital in Chapters 3, 4, and 5, and Hospital Episode Statistics covering the complete population of patients attending Type 1 A&E departments in England in Chapter 6. The scale of these data sets facilitated the examination of the entire affected population of interest in a real-life treatment setting, increasing the external validity of the results. The coverage of Hospital Episode Statistics also provides data on natural controls not affected by the service interventions of interest. As a result, all empirical evaluations presented have been able to use more robust analytical approaches including a comparator group.

8.3.2 Weaknesses

Whilst the two example service interventions examined in this thesis exhibit differing characteristics, they inevitably do not encompass the full range of service interventions which may be introduced in the NHS. Both examples studied were implemented in secondary care, and the evaluations presented considered emergency rather than elective patients. Interventions specifically targeting the healthcare workforce, such as changes to skill-mix, also form a major part of changes to the organisation and delivery of health services, but were not examined here. These service interventions may pose additional challenges not identified in this thesis.

Although there are many advantages of using Hospital Episode Statistics, a limitation of its use for economic evaluation is the lack of necessary preference-based health-related quality of life data needed to estimate QALYs. This was partially overcome through the development of the DANQALE tariff in Chapter 2, which involved estimating a discounted and quality-adjusted life expectancy tariff which could then be applied to estimated mortality differences associated with the programme under evaluation. However, this tariff estimates the change in QALYs associated with a change in mortality, and so does not capture any pure quality of life impacts not associated with mortality. It was therefore not possible to estimate any impacts purely on quality of life associated with either AQ or seven-day services as part of this thesis.

Given the positive impact of AQ both in terms of improvements in the quality of care provided and reductions in mortality associated with the programme, any pure quality of life impacts that the programme may have had are more likely to have been positive than negative. It is therefore unlikely that the ability to capture any such gains in quality of life would have changed the conclusions of Chapter 2. AQ was found to have likely represented a cost-effective use of resources, and so any increase in the magnitude of the benefits attributable to the programme would simply reinforce this conclusion.

Chapter 5 acknowledges that extending hospital services at weekends may have benefits beyond mortality, including improvements in quality of life. There is, therefore, a possibility that the programme could result in gains in quality of life not captured by the evaluation presented. However, Chapter 5 began by reviewing the evidence being used to support the case for seven-day services, highlighting the lack of evidence supporting the impact of service extensions on any patient outcomes. Given the absence of any evidence to suggest that extending hospital services at weekends would result in any patient benefit, and the potential opportunity costs in terms of resources diverted away from patients admitted to hospital during the week, there is no evidence to suggest that the planned reorganisation is cost-effective. Research into outcomes other than mortality would make a significant contribution to this area, which has to date focused on mortality differences.

A more general weakness associated with Hospital Episode Statistics is the absence of data on severity of illness (Bottle et al., 2014). All empirical analyses in this thesis controlled for numerous factors known to be associated with mortality risk, such as age, sex, and primary and secondary diagnoses. However, in common with other studies utilising administrative data, it was not possible to take account of the severity of these diagnoses. Recent papers examining the risks associated with weekend hospital admission have included a range of severity indicators, such condition-specific scales measuring stroke severity (Li and Rothwell, 2016). However, these severity indicators apply only to specific patient groups and are not collected in national administrative data sets. It was therefore not possible to include such severity indicators in the analysis presented in this thesis.

The difference-in-differences approach employed to evaluate AQ (Chapters 2, 3, and 4) reduces the risk that any unobserved severity differences could have biased the results presented. Variations in the severity of illness of the admitted populations between the North West and the rest of England would not impact on the estimates of the impact of AQ providing that these differences were fixed over time, as they would then be accounted for through the differencing procedure (Craig et al., 2012). A *differential change* in the severity of the admitted patient populations would therefore need to occur between the North West and the rest of England *after* the implementation of AQ for unobserved differences in severity to bias the results.

The importance of potential unobserved severity differences was examined in Chapter 6 in relation to estimates of the weekend effect in previous research used to support the seven-day services policy. As data on severity was unavailable, we examined the issue indirectly by examining Trusts' propensities to admit patients who attended A&E. This analysis concluded that Trusts appear to apply a more stringent admission threshold at weekends to patients seeking emergency care in A&E, raising the possibility that the population admitted at weekends is on average sicker than the population admitted on weekdays in ways not completely captured by standard risk-adjustment. Chapter 6 highlighted how this increased threshold for admission at weekends is likely to have biased previous studies on weekend mortality. Expanding the analysis performed in Chapter 6 to include all patients attending A&E eliminates any biases introduced by severity differences amongst the admitted patient population. However, the possibility remains that the populations of patients attending A&E on weekdays compared to weekends differ in terms of unobserved severity. A subsequent study utilising arrival by ambulance as an indicator of patient severity confirmed that patients admitted to hospital at weekends through A&E do appear to be more severe than their counterparts admitted on weekdays (Anselmi et al., 2016), supporting the conclusions of Chapter 6.

Finally, there may be concerns about data quality. The Hospital Episode Statistics data utilised in this thesis is derived from patient records, and is designed for secondary analytical use (NHS Digital, 2012). The data are validated monthly by the Health and Social Care Information Centre (HSCIC), and data quality reports and checks are completed before the data are made available. Information on these quality checks is published, and the relevant quality reports were reviewed for each chapter. No quality issues were identified with the data utilised in any of the empirical analyses presented.

Whilst the quasi-experimental introduction of the AQ programme to just one region of England increased the external validity of the evaluations presented in Chapters 2, 3, and 4, natural experimental approaches are more susceptible to bias and confounding than randomised studies (Craig et al., 2012; Gillies et al., 2016). The difference-in-differences approach taken assumes that in the absence of treatment, the average outcomes for the treated and control groups examined would have followed the same trends over time (Abadie, 2005). Difference-in-differences is considered to be one of the most robust analytical approaches available in the absence of treatment randomisation as it controls for selection bias originating from both observable and unobservable characteristics, providing that these unobservable characteristics are fixed over time (Craig et al., 2012).

Selection into the AQ programme was on a regional basis, with universal participation of Trusts in the North West region of England. Universal participation reduced the risk of bias as treatment assignment was not based on individual provider characteristics. This provided us with a stronger evaluation design than for AQ’s predecessor in the United States of America (USA), the Premier Hospital Quality Incentive Demonstration (HQID), where hospitals self-selected into the programme (Jha et al., 2012).

To maximise the internal validity of the analyses performed, the Medical Research Council (MRC) guidelines for using natural experiments to evaluate population health interventions were followed throughout (Craig et al., 2012). These recommendations included the use of multiple pre and post observation points, and utilisation of accurate data on exposure to the programme, potential confounders, and outcomes. The parallel trends assumption was also verified, that is, that although treatment and control groups may have different levels of the outcome of interested prior to the introduction of treatment, their trends in pre-introduction outcomes should be the same (Ryan et al., 2015).

Ryan and colleagues provide a difference-in-differences checklist as a guide to performing high quality studies using this technique (Ryan et al., 2015). The analyses conducted in this thesis meet all of the criteria listed in this checklist. The only exception is in relation to criterion number seven, “treatment does not ‘spill-over’ from treatment group to comparison group”. Some evidence of a positive spillover effect was detected in the longer-term evaluation of AQ presented in Chapter 3. The impact of these potential spillover effects was therefore explored in two sensitivity analyses, and the existence of spillovers suggested as one explanation for the lack of impact of AQ on mortality over the longer-term.

There have been significant advancements and refinements in the methods for observational evaluation in the absence of treatment randomisation over recent years (Gillies et al., 2016). Many of these have focused on solutions to overcome non-parallel trends in pre-intervention outcomes.

The AQ programme was used as an example upon which to compare the performance of difference-in-differences to that of the synthetic control method, which constructs a weighted combination of control units based on their pre-intervention outcomes and relaxes the parallel trends assumption (Kreif et al., 2016). In contrast with the original difference-in-differences analysis of the AQ programme (Sutton et al., 2012), and that presented in Chapter 2 of this thesis, the synthetic control method reported that AQ did not significantly reduce mortality during its first 18 months.

However, a later simulation study comparing the performance of difference-in-differences, synthetic control, and two alternative methods, confirmed the superiority of difference-in-differences when the parallel trends assumption is met (O'Neill et al., 2016), which appeared to be the case for the AQ evaluations presented in this thesis. In situations where the parallel trends assumption holds, it was demonstrated that difference-in-differences produces unbiased estimates of the true treatment effect, whilst the synthetic control method produced the most biased and inefficient estimates. Although the synthetic control approach can mitigate some of the bias present in difference-in-differences estimates when the parallel trends assumption is violated, synthetic control still produced inefficient estimates compared to a lagged dependent variable approach (O'Neill et al., 2016). This further work confirms that difference-in-differences was therefore the correct approach to take in the evaluation of AQ, given that the parallel trends assumption appeared not to be violated in this case.

8.4 Reflecting across the thesis

The chapters of this thesis are interconnected, and the findings of each study therefore have implications for the other chapters. Chapter 2 presented an initial assessment of the cost-effectiveness of the AQ programme, with the effect of the scheme evaluated in terms of its impact on mortality within 30 days of admission to hospital. These effects on short-term mortality were then converted to projected gains in QALYs using published estimates of life expectancy for the general population, adjusted for quality of life, again using mean EQ-5D values from a general population sample. It was acknowledged that the use of this age-sex specific DANQALE tariff was likely to overestimate the health gains enjoyed by the additional survivors because the average life expectancy and health-related quality of life of individuals admitted to hospital for acute myocardial infarction (AMI), heart failure, and pneumonia are likely to be lower than that of the general population.

In Chapter 4, attempts were made to overcome some of these limitations using survival analysis to improve the accuracy of estimated life year gains attributable to AQ. Utilising survival models resulted in a reduction in the estimated remaining life expectancy of the patient cohort both in the presence and absence of the AQ programme. These lower absolute estimates of remaining life expectancy were anticipated as this method utilised information on the mortality rates observed amongst the patient population rather than general population life expectancy estimates. Nevertheless, employing this method resulted in a larger estimated treatment effect attributable to

AQ in comparison with the approach taken in the original cost-effectiveness analysis. This suggested that AQ had a prolonged impact on survival past the 30-day post-admission window previously assessed in Chapter 2.

In this instance, the increased magnitude of the treatment effect associated with AQ would not have changed the conclusions of Chapter 2. AQ was found to have likely represented a cost-effective use of resources, and so an increase in the magnitude of the benefits attributable to the programme simply reinforces this conclusion.

Chapter 5 also utilised the DANQALE tariff developed in Chapter 2 to estimate the potential QALY gain associated with eradicating the weekend effect, based on the differences in mortality within 30 days of emergency hospital admission. Mortality was assessed over a period of 30 days for consistency with the existing literature in this area, and in particular the research used to support the seven-day services policy. Utilising survival analysis to compare the survival of patients admitted at weekends compared to weekdays may have again resulted in different estimates of the remaining life expectancy of the cohorts, and in turn altered the magnitude of the potential QALY gain attainable if the weekend effect were to be eradicated.

However, subsequent analysis conducted in Chapter 6 showed that the weekend effect was driven by a reduction in the volumes of patients admitted to hospital at the weekend rather than an increase in the number of deaths. The seven-day services policy is only likely to be successful in reducing mortality if reduced availability of services in hospitals on the day of admission is the major cause of the weekend effect. The results presented in Chapter 6 suggest that this is not the case, indicating that the planned service extensions are unlikely to reduce the absolute number of deaths. Consequently, these results suggest that the potential QALY gain achievable as a result of implementing the seven-day services policy is likely to be significantly less than that estimated in Chapter 5. This finding only further reinforces the conclusions of Chapter 5, that there is as yet no clear evidence: that seven day working will, in isolation, reduce the weekend death rate, that lower weekend mortality rates can be achieved without increasing weekday death rates; or that such reorganisation is cost-effective.

The findings of Chapter 6 could also have potential implications for Chapters 2, 3, and 4, which all focus on comparisons of mortality rates across populations of patients admitted to hospital in an emergency. Chapter 6 showed that variations in mortality rates on different days of the week can be explained by variations in admission thresholds not previously captured by standard risk-adjustment methods. This raises the possibility that variations in mortality rates across hospitals may also reflect variations in admission thresholds.

The evaluations of AQ (Chapters 2, 3, and 4) all rely on comparisons of mortality rates amongst patients admitted to hospitals in the North West with mortality rates of patients admitted to hospitals in the rest of England. Due to the difference-in-differences analytical approach used, variations in hospitals' admission thresholds between the North West and the rest of England would not impact upon the results presented, as these differences are accounted for through the differencing procedure (Craig et al., 2012). A key strength of the difference-in-differences approach is that it controls for differences in both observable and unobservable characteristics between the two groups being compared, providing that these characteristics are time-invariant. A *differential change* in the admission thresholds of hospitals would therefore have needed to occur between the North West and the rest of England *after* the implementation of AQ for differences in admission thresholds to bias the results.

8.5 Implications for policy and future research

The work presented in this thesis demonstrates that whilst service interventions are inherently different to pharmaceuticals in some important respects, they are still amenable to economic evaluation.

8.5.1 Implications for policymakers

Subjecting service interventions to systematic and rigorous cost-effectiveness evaluation, as is now mandated for all health technologies, would ensure parity in funding decisions between health technologies and service interventions. This would also increase transparency in resource allocation decisions, and improve allocative efficiency if service interventions failing to demonstrate cost-effectiveness were not introduced as a result. Economic evaluation of service interventions could be undertaken by NICE, or another similar agency. The evaluation of seven-day services presented in Chapter 5 gives some indication of the scale of resources currently being committed without supportive evaluative evidence. It is estimated that this one policy alone will cost 1.5% to 2% of hospital income, equivalent to a 5% to 6% increase in the annual costs of emergency admissions.

The seven-day services policy perfectly illustrates the dangers of relying on cross-sectional associations, and confusing correlation with causation. As yet there is no evidence to support the planned intervention or to suggest that it will result in any benefit to patients. Whilst extending services has the potential to have benefits to dimensions other than mortality, evidence to demonstrate such benefits should be required before the programme is introduced.

The AQ P4P programme was found to represent a cost-effective use of resources during its first 18 months of operation. The fact that the programme had a more positive effect than many previous P4P initiatives has been attributed to the regional and collaborative nature of the scheme in the short-term. The supporting mechanisms provided alongside the financial incentives appeared to be key, suggesting that additional accompanying levers can significantly enhance the effectiveness of P4P programmes.

Whilst the conclusions around the benefits of AQ during its first 18 months are clear, the value of the programme over the longer-term is uncertain. The optimal lifetime of P4P schemes and the impact of their removal remains unknown. The effect of such removal will depend upon whether quality improvement is a transitory or investment activity. This question could be answered if policymakers were willing to experiment with discontinuing incentive programmes whilst continuing to collect data on performance against the previously incentivised quality measures.

8.5.2 Implications for future research

In order for quality economic evaluations of service interventions to be conducted, methodological improvements must still be made. In particular, developments are necessary for estimating the impact of service evaluations on costs and QALYs. Routine collection of preference-based health-related quality of life measures such as the EQ-5D would facilitate greater economic evaluation of service interventions, potentially through an expansion of the national patient-reported outcome measures (PROMs) programme. However, collection of baseline measurement in patients utilising emergency care poses significant difficulties, as these patients will have had no reason to complete such an assessment (Gibbons et al., 2016). Alternative methods for quantifying the effects of service interventions in terms of QALYs in the absence of primary data on health-related quality of life, such as the DANQALE method developed in this thesis, are therefore still likely to be required. Further development of such tariffs would allow economic evaluations to be performed at far lower cost, and with minimal patient burden, compared to primary data collection from all affected patients.

A wide range of methodological approaches may be needed to overcome the challenges faced when conducting economic evaluations of service interventions. For example, this thesis has shown that utilising survival analysis techniques commonly employed in clinical trials can improve treatment effect estimates associated with service interventions. The value of ex ante analysis has also been demonstrated, and the importance of considering potential spillover effects.

The potential presence of spillover effects, and how these were treated, was crucial for the interpretation of the AQ evaluation results. Although the performance of Trusts in the AQ region continued to improve over the longer-term, both in terms of the process measures of care and mortality rates, mortality declined at a faster rate in the rest of England. It was therefore not possible to determine whether this finding indicated that the incentives were not effective over this longer-term period, or if this pattern of results represented positive spillover effects of the programme. Fully understanding these results and the longer-term impact of service interventions in general requires some complex issues to be resolved. A structured method for identifying and measuring spillover effects is needed if the true impacts of service interventions are to be correctly quantified and understood.

Further research into the population of patients directly admitted to hospital in an emergency is required as relatively little is known about this group. The daily patterns of admission volumes and mortality rates for this patient group were markedly different to those admitted through A&E, and future research should look to understand the reasons for this.

Differences in Trusts’ propensities to admit patients on weekdays compared to weekends was found to explain the previously detected weekend effect in mortality. Future research should examine the reasons why Trusts’ propensities to admit patients differ by day of the week, and what drives admission decisions.

Finally, the findings of Chapter 6 have potential implications for all studies comparing mortality rates across admitted patient populations. Variations in mortality rates for patients admitted on different days of the week were found to be explained by variations in admission thresholds not previously captured by standard risk-adjustment methods. It is therefore possible that this finding could extend to differences in admission thresholds across Trusts. Examining Trusts’ propensities to admit patients attending A&E offers a partial solution to this issue, but further research is needed to fully understand the influence of admission propensities on mortality rates amongst the admitted population.

References

Abadie, A., 2005. Semiparametric Difference-in-Differences Estimators. Rev. Econ. Stud. 72, 1–19.

Academy of Medical Royal Colleges, 2012. Seven Day Consultant Present Care. 2012, London: Academy of Medical Royal Colleges.

Allen, T., Fichera, E., Sutton, M., 2016. Can Payers Use Prices to Improve Quality? Evidence from English Hospitals. Health Econ. 25, 56–70.

An, L.C., Bluhm, J.H., Foldes, S.S., Alesci, N.L., Klatt, C.M., Center, B.A., Nersesian, W.S., Larson, M.E., Ahluwalia, J.S., Manley, M.W., 2008. A randomized trial of a pay-for-performance program targeting clinician referral to a state tobacco quitline. Arch. Intern. Med. 168, 1993–1999. doi:10.1001/archinte.168.18.1993

Anderson, R., Hardwick, R., 2016. Realism and resources: Towards more explanatory economic evaluation. Evaluation 22, 323–341. doi:10.1177/1356389016652742

Angrist, J.D., Pischke, J.-S., 2008. Mostly Harmless Econometrics: An Empiricist's Companion. Princeton University Press.

Anselmi, L., Meacock, R., Kristensen, S.R., Doran, T., Sutton, M., 2016. Arrival by ambulance explains variation in mortality by time of admission: retrospective study of admissions to hospital following emergency department attendance in England. BMJ Qual Saf. doi:10.1136/bmjqs-2016-005680

Appleby, J., 2016. A 7/7 NHS: what price equity? BMJ 352, i404. doi:10.1136/bmj.i404

Ashenfelter, O., 1978. Estimating the effect of training programs on earnings. Rev. Econ. Stat. 60, 47–57.

Australian Government Department of Health, 2015. Guidelines for Preparing Submissions to the Pharmaceutical Benefits Advisory Committee Version 4.5. Commonwealth of Australia, Canberra, Australia.

Aylin, P., 2015. Making sense of the evidence for the "weekend effect." The BMJ 351, h4652. doi:10.1136/bmj.h4652

Aylin, P., Yunus, A., Bottle, A., Majeed, A., Bell, D., 2010. Weekend mortality for emergency admissions. A large, multicentre study. Qual. Saf. Health Care 19, 213–217. doi:10.1136/qshc.2008.028639

Bagust, A., Beale, S., 2013. Survival Analysis and Extrapolation Modeling of Time-to-Event Clinical Trial Data for Economic Evaluation An Alternative Approach. Med. Decis. Making. doi:10.1177/0272989X13497998

Barratt, H., Campbell, M., Moore, L., Zwarenstein, M., Bower, P., 2016. Randomised controlled trials of complex interventions and large-scale transformation of services. NIHR Journals Library.

Bell, C.M., Redelmeier, D.A., 2001. Mortality among Patients Admitted to Hospitals on Weekends as Compared with Weekdays. N. Engl. J. Med. 345, 663–668.

Bottle, R., Gaudoin, R., Goudie, R., Jones, S., Aylin, P., 2014. Can valid and practical risk-prediction or casemix adjustments models, including adjustment for comorbidity, be generated from

English hospital administrative data (Hospital Episode Statistics)? A national observational study. Health Serv. Deliv. Res. 2. doi:10.3310/hsdr02400

Brazier, J.E., Yang, Y., Tsuchiya, A., Rowen, D.L., 2010. A review of studies mapping (or cross walking) non-preference based measures of health to generic preference-based measures. Eur. J. Health Econ. HEPAC Health Econ. Prev. Care 11, 215–225.

Campbell, S.M., Reeves, D., Kontopantelis, E., Sibbald, B., Roland, M., 2009. Effects of Pay for Performance on the Quality of Primary Care in England. N. Engl. J. Med. 361, 368–378.

Canadian Agency for Drugs and Technologies in Health, 2006. Guidelines for the Economic Evaluation of Health Technologies: Canada ’3rd Edition].

Castelli, A., Dawson, D., Gravelle, H., Jacobs, R., Kind, P., Loveridge, P., Martin, S., O’Mahony, M., Stevens, P.A., Stokes, L., Street, A., Weale, M., 2007. A new approach to measuring health system output and Productivity. Natl. Inst. Econ. Rev. 200, 105–117. doi:10.1177/0027950107080395

Chalkidou, K., Culyer, A., Naidoo, B., Littlejohns, P., 2008. Cost-effective public health guidance: asking questions from the decision-maker’s viewpoint. Health Econ. 17, 441–448. doi:10.1002/hec.1277

Ciani, O., Wilcher, B., van Giessen, A., Taylor, R.S., 2017. Linking the Regulatory and Reimbursement Processes for Medical Devices: The Need for Integrated Assessments. Health Econ. 26, 13–29. doi:10.1002/hec.3479

Claxton, K., Martin, S., Soares, M., Rice, N., Spackman, E., Hinde, S., Devlin, N., Smith, P., Sculpher, M., 2015. Methods for the estimation of the National Institute for Health and Care Excellence cost-effectiveness threshold. Health Technol. Assess. 19.

Claxton, K., Martin, S., Soares, M., Rice, N., Spackman, E., Hinde, S., Devlin, N., Smith, P., Sculpher, M., 2013. Methods for the Estimation of the NICE Cost Effectiveness Threshold (No. 81), CHE Research Papers. The University of York, York.

Claxton, K., Sculpher, M., Drummond, M., 2002. A rational framework for decision making by the National Institute For Clinical Excellence (NICE). The Lancet 360, 711–715. doi:10.1016/S0140-6736(02)09832-X

Coast, J., 2004. Is economic evaluation in touch with society’s health values? BMJ 329, 1233. doi:10.1136/bmj.329.7476.1233

Congressional Budget Office, 2007. CBO’s Health Insurance Simulation Model: A Technical Description.

Craig, P., Cooper, C., Gunnell, D., Haw, S., Lawson, K., Macintyre, S., Ogilvie, D., Petticrew, M., Reeves, B., Sutton, M., Thompson, S., 2012. Using natural experiments to evaluate population health interventions: new Medical Research Council guidance [WWW Document]. URL http://jech.bmj.com/content/66/12/1182.long (accessed 5.3.16).

Craig, P., Dieppe, P., Macintyre, S., Michie, S., Nazareth, I., Petticrew, M., 2008. Developing and evaluating complex interventions: the new Medical Research Council guidance. BMJ 337, a1655. doi:10.1136/bmj.a1655

Craig, P., Katikireddi, S., Leyland, A., Popham, F., 2017. Natural Experiments: An Overview of Methods, Approaches, and Contributions to Public Health Intervention Research. Annu. Rev. Public Health 38, null. doi:10.1146/annurev-publhealth-031816-044327

Crump, H., 2015. Seven day working: why the health secretary's proposal is not as simple as it sounds. The BMJ 351, h4473. doi:10.1136/bmj.h4473

Curtin, K., Beckman, H., Pankow, G., Milillo, Y., Greene, R.A., 2006. Return on Investment in Pay for Performance: A Diabetes Case Study. J. Healthc. Manag. 51, 365–376.

Davies, C., Briggs, A., Lorgelly, P., Garellic, G., Malchau, H., 2013. The “Hazards” of Extrapolating Survival Curves. Med. Decis. Making 0272989X12475091. doi:10.1177/0272989X12475091

Dawson, D., Gravelle, H., O'Mahony, M., Steet, A., Weale, M., Castelli, A., Jacobs, R., Kind, P., Loveridge, P., Martin, S., Stevens, P., Stokes, L., 2005. Developing New Approaches to Measuring NHS Outputs and Activity (No. 6), CHE Research Papers. The University of York.

Department for Communities and Local Government, 2011. English indices of deprivation 2010 [WWW Document]. URL https://www.gov.uk/government/statistics/english-indices-of-deprivation-2010 (accessed 10.13.15).

Department of Health, 2015. 7-day NHS services: a factsheet - GOV.UK [WWW Document]. URL https://www.gov.uk/government/publications/7-day-nhs-services-a-factsheet/7-day-nhs-services-a-factsheet (accessed 8.6.15).

Department of Health, 2014. Annual Report and Accounts 2013-14.

Department of Health, 2010. Using the Commissioning for Quality and Innovation (CQUIN) payment framework - A summary guide.

Department of Health, 2008. Guidance on the routine collection of Patient Reported Outcome Measures (PROMs).

Doran, T., Bloor, K., Maynard, A., 2015. The death of death rates? BMJ 351, h3466. doi:10.1136/bmj.h3466

Doran, T., Kontopantelis, E., Valderas, J.M., Campbell, S., Roland, M., Salisbury, C., Reeves, D., 2011. Effect of financial incentives on incentivised and non-incentivised clinical activities: longitudinal analysis of data from the UK Quality and Outcomes Framework. BMJ 342, d3590–d3590. doi:10.1136/bmj.d3590

Drummond, M., Griffin, A., Tarricone, R., 2009. Economic Evaluation for Devices and Drugs—Same or Different? Value Health 12, 402–404. doi:10.1111/j.1524-4733.2008.00476_1.x

Drummond, M., Sculpher, M., Claxton, K., Stoddart, G., Torrance, G., 2015. Methods for the Economic Evaluation of Health Care Programmes Fourth Edition. Oxford University Press, Incorporated.

Drummond, M.F., Jefferson, T.O., 1996. Guidelines for authors and peer reviewers of economic submissions to the BMJ. The BMJ Economic Evaluation Working Party. BMJ 313, 275–283.

Eijkenaar, F., 2012. Pay for Performance in Health Care An International Overview of Initiatives. Med. Care Res. Rev. 69, 251–276. doi:10.1177/1077558711432891

Eijkenaar, F., Emmert, M., Scheppach, M., Schöffski, O., 2013. Effects of pay for performance in health care: A systematic review of systematic reviews. Health Policy 110, 115–130. doi:10.1016/j.healthpol.2013.01.008

Emmert, M., Eijkenaar, F., Kemter, H., Esslinger, A.S., Schöffski, O., 2011. Economic evaluation of pay-for-performance in health care: a systematic review. Eur. J. Health Econ. doi:10.1007/s10198-011-0329-8

Finkelstein, A., Taubman, S., 2015. Randomize evaluations to improve health care delivery. Science 347, 720–722. doi:10.1126/science.aaa2362

Franken, M., Le Polain, M., Cleemput, I., Koopmanchap, M., 2012. Similarities and differences between five European drug reimbursement ssytems 28, 349–357.

Freemantle, N., Ray, D., McNulty, D., Rosser, D., Bennett, S., Keogh, B.E., Pagano, D., 2015. Increased mortality associated with weekend hospital admission: a case for expanded seven day services? The BMJ 351, h4596. doi:10.1136/bmj.h4596

Freemantle, N., Richardson, M., Wood, J., Ray, D., Khosla, S., Shahian, D., Roche, W.R., Stephens, I., Keogh, B., Pagano, D., 2012. Weekend hospitalization and additional risk of death: an analysis of inpatient data. J. R. Soc. Med. 105, 74–84. doi:10.1258/jrsm.2012.120009

Frew, E., 2017. Aligning Health Economics Methods to Fit with the Changing World of Public Health. Appl. Health Econ. Health Policy. 1–3. doi:10.1007/s40258-017-0319-9

Gelber, R.D., Goldhirsch, A., Cole, B.F., 1993. Parametric extrapolation of survival estimates with applications to quality of life evaluation of treatments. International Breast Cancer Study Group. Control. Clin. Trials 14, 485–499.

Gibbons, E., Black, N., Fallowfield, L., Newhouse, R., Fitzpatrick, R., 2016. Patient-reported outcome measures and the evaluation of services. NIHR Journals Library.

Gillies, C., Freemantle, N., Grieve, R., Sekhon, J., Forder, J., 2016. Advancing quantitative methods for the evaluation of complex interventions. NIHR Journals Library.

Glickman, S.W., Ou, F.-S., DeLong, E.R., Roe, M.T., Lytle, B.L., Mulgund, J., Rumsfeld, J.S., Gibler, W.B., Ohman, E.M., Schulman, K.A., Peterson, E.D., 2007. Pay for performance, quality of care, and outcomes in acute myocardial infarction. JAMA J. Am. Med. Assoc. 297, 2373–2380. doi:10.1001/jama.297.21.2373

Godlee, F., 2015. What to do about the “weekend effect.” BMJ 351, h4840. doi:10.1136/bmj.h4840

Gold, M.R., 1996. Cost-Effectiveness in Health and Medicine. Oxford University Press.

Gomes, M., Gutacker, N., Bojke, C., Street, A., 2016. Addressing Missing Data in Patient-Reported Outcome Measures (PROMS): Implications for the Use of PROMS for Comparing Provider Performance. Health Econ. 25, 515–528. doi:10.1002/hec.3173

Graves, N., Walker, D., Raine, R., Hutchings, A., Roberts, J.A., 2002. Cost data for individual patients included in clinical studies: no amount of statistical analysis can compensate for inadequate costing methods. Health Econ. 11, 735–739. doi:10.1002/hec.683

Greene, S.E., Nash, D.B., 2009. Pay for Performance: An Overview of the Literature. Am. J. Med. Qual. 24, 140–163. doi:10.1177/1062860608326517

Grieve, R., Hawkins, N., Pennington, M., 2013. Extrapolation of Survival Data in Cost-effectiveness Analyses Improving the Current State of Play. Med. Decis. Making 33, 740–742. doi:10.1177/0272989X13492018

Griffin, S., Rice, N., Sculpher, M., 2009. Economic evaluation of public health interventions, in: Evidence-Based Public Health: Effectiveness and Efficiency.

Grossbart, S.R., 2006. What's the return? Assessing the effect of "pay-for-performance" initiatives on the quality of care delivery. Med. Care Res. Rev. MCRR 63, 29S–48S.

Gutacker, N., Bloor, K., Cookson, R., 2015. Comparing the performance of the Charlson/Deyo and Elixhauser comorbidity measures across five European countries and three conditions. Eur. J. Public Health 25, 15–20. doi:10.1093/eurpub/cku221

Haynes, L., Service, O., Goldacre, B., Torgerson, D., 2012. Test, Learn, Adapt: Developing Public Policy with Randomised Controlled Trials [WWW Document]. URL https://www.gov.uk/government/publications/test-learn-adapt-developing-public-policy-with-randomised-controlled-trials (accessed 3.31.16).

HCUP, 2009. Classifications Software (CCS) for Mortality Reporting [WWW Document]. URL http://hcup-us.ahrq.gov/toolssoftware/icd_10/ccs_icd_10.jsp (accessed 12.2.15).

Health and Social Care Information Centre, 2015. Summary Hospital-level Mortality Indicator (SHMI) - Frequently Asked Questions (FAQs).

Health and Social Care Information Centre, 2013. Linked HES-ONS mortality data [WWW Document]. URL http://www.hscic.gov.uk/article/2677/Linked-HES-ONS-mortality-data (accessed 9.29.14).

Health and Social Care Information Centre, 2012. Hospital Episode Statistics [WWW Document]. URL http://www.hscic.gov.uk/hes (accessed 9.29.14).

Heckman, J., 2008. Econometric Causality. Int. Stat. Rev. 76, 1–27.

Hogan, H., Zipfel, R., Neuburger, J., Hutchings, A., Darzi, A., Black, N., 2015. Avoidability of hospital deaths and association with hospital-wide mortality ratios: retrospective case record review and regression analysis. BMJ 351, h3239. doi:10.1136/bmj.h3239

Holmstrom, B., Milgrom, P., 1991. Multitask Principal-Agent Analyses: Incentive Contracts, Asset Ownership, and Job Design. J. Law Econ. Organ. 7, 24.

House of Commons Library, 2015. Research Briefings - Accident and Emergency Statistics [WWW Document]. URL http://researchbriefings.parliament.uk/ResearchBriefing/Summary/SN06964#fullreport (accessed 8.6.15).

Hunter, R.M., Davie, C., Rudd, A., Thompson, A., Walker, H., Thomson, N., Mountford, J., Schwamm, L., Deanfield, J., Thompson, K., Dewan, B., Mistry, M., Quoraishi, S., Morris, S., 2013. Impact on Clinical and Cost Outcomes of a Centralized Approach to Acute Stroke Care in London: A Comparative Effectiveness Before and After Model. PLOS ONE 8, e70420. doi:10.1371/journal.pone.0070420

Jha, A.K., Joynt, K.E., Orav, E.J., Epstein, A.M., 2012. The Long-Term Effect of Premier Pay for Performance on Patient Outcomes. N. Engl. J. Med. 366, 1606–1615. doi:10.1056/NEJMsa1112351

Jones, A., Rice, N., 2011. Econometric Evaluation of Health Policies, in: The Oxford Handbook of Health Economics. Oxford University Press, Oxford.

Kaarboe, O., Siciliani, L., 2011. Multi-tasking, quality and pay for performance. Health Econ. 20, 225–238. doi:10.1002/hec.1582

Kelman, S., Friedman, J.N., 2009. Performance Improvement and Performance Dysfunction: An Empirical Examination of Distortionary Impacts of the Emergency Room Wait-Time Target in the English National Health Service. J. Public Adm. Res. Theory 19, 917–946. doi:10.1093/jopart/mun028

Knapp, M., Beecham, J., Anderson, J., Dayson, D., Leff, J., Margolius, O., O'Driscoll, C., Wills, W., 1990. The TAPS project. 3: Predicting the community costs of closing psychiatric hospitals. Br. J. Psychiatry 157, 661–670. doi:10.1192/bjp.157.5.661

Kouides, R.W., Bennett, N.M., Lewis, B., Cappuccio, J.D., Barker, W.H., LaForce, F.M., 1998. Performance-based physician reimbursement and influenza immunization rates in the elderly. The Primary-Care Physicians of Monroe County. Am. J. Prev. Med. 14, 89–95.

Kreif, N., Grieve, R., Hangartner, D., Turner, A., Nikolova, S., Sutton, M., 2016. Examination of the Synthetic Control Method for Evaluating Health Policies with Multiple Treated Units. Health Econ. 25, 1514–1528.

Kristensen, S., Meacock, R., Sutton, M., 2015. Methodology for Modelling Changes in Health Services Delivery and Policy. Vignette prepared for the MRC-NIHR Methodology Research Programme Advisory Group.

Kristensen, S.R., McDonald, R., Sutton, M., 2013. Should pay-for-performance schemes be locally designed? evidence from the commissioning for quality and innovation (CQUIN) framework. J. Health Serv. Res. Policy 18, 38–49. doi:10.1177/1355819613490148

Lamont, T., Barber, N., Pury, J. de, Fulop, N., Garfield-Birkbeck, S., Lilford, R., Mear, L., Raine, R., Fitzpatrick, R., 2016. New approaches to evaluating complex health and care systems. BMJ 352, i154. doi:10.1136/bmj.i154

Latimer, N.R., 2013. Survival analysis for economic evaluations alongside clinical trials--extrapolation with patient-level data: inconsistencies, limitations, and a practical guide. Med. Decis. Mak. Int. J. Soc. Med. Decis. Mak. 33, 743–754. doi:10.1177/0272989X12472398

Lee, T.-T., Cheng, S.-H., Chen, C.-C., Lai, M.-S., 2010. A pay-for-performance program for diabetes care in Taiwan: a preliminary assessment. Am. J. Manag. Care 16, 65–69.

Lester, H., Schmittdiel, J., Selby, J., Fireman, B., Campbell, S., Lee, J., Whippy, A., Madvig, P., 2010. The impact of removing financial incentives from clinical quality indicators: longitudinal analysis of four Kaiser Permanente indicators. BMJ 340, c1898–c1898. doi:10.1136/bmj.c1898

Li, L., Rothwell, P., 2016. Biases in detection of apparent "weekend effect" on outcome with administrative coding data: population based study of stroke. BMJ 353, 2648. doi:10.1136/bmj.i2648

Lilford, R., Mohammed, M.A., Spiegelhalter, D., Thomson, R., 2004. Use and misuse of process and outcome data in managing performance of acute medical care: avoiding institutional stigma. Lancet Lond. Engl. 363, 1147–1154. doi:10.1016/S0140-6736(04)15901-1

Lilford, R.J., Chen, Y.-F., 2015. The ubiquitous weekend effect: moving past proving it exists to clarifying what causes it. BMJ Qual. Saf. bmjqs-2015-004360. doi:10.1136/bmjqs-2015-004360

Lilford, R.J., Chilton, P.J., Hemming, K., Girling, A.J., Taylor, C.A., Barach, P., 2010. Evaluating policy and service interventions: framework to guide selection and interpretation of study end points. BMJ 341, c4413. doi:10.1136/bmj.c4413

Lindenauer, P.K., Remus, D., Roman, S., Rothberg, M.B., Benjamin, E.M., Ma, A., Bratzler, D.W., 2007. Public Reporting and Pay for Performance in Hospital Quality Improvement. N. Engl. J. Med. 356, 486–496. doi:10.1056/NEJMsa064964

Malani, A., Reif, J., 2015. Interpreting pre-trends as anticipation: Impact on estimated treatment effects from tort reform. J. Public Econ. 124, 1–17. doi:10.1016/j.jpubeco.2015.01.001

Mason, A., Walker, S., Claxton, K., Cookson, R., Fenwick, E., Sculpher, M., 2008. The GMS Quality and Outcoms Framework: Are the Quality and Outcomes Framework (QOF) Indicators a Cost-Effective Use of NHS Resources?, Quality and Outcomes Framework Joint Executive Summary: Reports to the Department of Health from the University of East Anglia & The University of York.

Maynard, A., 2012. The powers and pitfalls of payment for performance. Health Econ. 21, 3–12. doi:10.1002/hec.1810

McCartney, M., 2015. Margaret McCartney: The zombie statistic behind the push for seven day working. The BMJ 351, h3575. doi:10.1136/bmj.h3575

McDonald, R., Boaden, R., Roland, M., Kristensen, S.R., Meacock, R., Lau, Y.-S., Mason, T., Turner, A.J., Sutton, M., 2015. A qualitative and quantitative evaluation of the Advancing Quality pay-for-performance programme in the NHS North West. Health Serv. Deliv. Res. 3, 1–104. doi:10.3310/hsdr03230

McKee, M., 2015. Is the UK government right that seven day working in hospitals would save 6000 lives a year? The BMJ 351, h4723. doi:10.1136/bmj.h4723

Meacock, R., Doran, T., Sutton, M., 2015. What are the Costs and Benefits of Providing Comprehensive Seven-day Services for Emergency Hospital Admissions? Health Econ. 24, 907–912. doi:10.1002/hec.3207

Meacock, R., Kristensen, S.R., Sutton, M., 2014. The Cost-Effectiveness of Using Financial Incentives to Improve Provider Quality: A Framework and Application. Health Econ. 23, 1–13. doi:10.1002/hec.2978

Medical Research Council, 2009. Using natural experiments to evaluate population health interventions: guidance for producers and users of evidence.

Medical Research Council, 2008. Developing and evaluating complex interventions: new guidance.

Mehrotra, A., Damberg, C.L., Sorbero, M.E.S., Teleki, S.S., 2009. Pay for Performance in the Hospital Setting: What Is the State of the Evidence? Am. J. Med. Qual. 24, 19–28. doi:10.1177/1062860608326634

Mohammed, M.A., Deeks, J.J., Girling, A., Rudge, G., Carmalt, M., Stevens, A.J., Lilford, R.J., 2009. Evidence of methodological bias in hospital standardised mortality ratios: retrospective database study of English hospitals. BMJ 338, b780. doi:10.1136/bmj.b780

Morris, S., Hunter, R., Ramsay, A., Boaden, R., McKevitt, C., Perry, C., Pursani, N., Rudd, A., Schwamm, L., Turner, A., Tyrrell, P., Wolfe, C., Fulop, N., 2014. Impact of centralising acute stroke services in English metropolitan areas on mortality and length of hospital stay: difference-in-differences analysis. BMJ 349, g4757. doi:10.1136/bmj.g4757

Mullen, K.J., Frank, R.G., Rosenthal, M.B., 2010. Can you get what you pay for? Pay-for-performance and the quality of healthcare providers. Rand J. Econ. 41, 64–91.

Nahra, T.A., Reiter, K.L., Hirth, R.A., Shermer, J.E., Wheeler, J.R.C., 2006. Cost-Effectiveness of Hospital Pay-for-Performance Incentives. Med. Care Res. Rev. 63, 49S–72S. doi:10.1177/1077558705283629

National Audit Office, 2013. Emergency admissions to hospital: managing the demand.

National Institute for Health and Clinical Excellence (NICE), 2013. Guide to the methods of technology appraisal 2013.

NHS Commissioning Board, 2013. Everybody Counts: Planning for Patients 2013/14. London: NHS Commissioning Board.

NHS Digital, 1 Trevelyan Square, 2012. Hospital Episode Statistics [WWW Document]. URL http://content.digital.nhs.uk/hes (accessed 4.11.17).

NHS England, 2017. Seven Day Hospital Services [WWW Document]. URL https://www.england.nhs.uk/ourwork/qual-clin-lead/seven-day-hospital-services/ (accessed 3.20.17).

NHS England, 2016. Seven Day Hospital Services - Our ambition [WWW Document]. URL https://www.england.nhs.uk/ourwork/qual-clin-lead/seven-day-hospital-services/our-ambition/ (accessed 1.27.17).

NHS England, 2014. The Forward View into action: Planning for 2015/16.

NHS England, Seven Days a Week Forum, 2013a. Summary of Initial Findings.

NHS England, Seven Days a Week Forum, 2013b. Evidence base and clinical standards for the care and onward transfer of acute inpatients.

NHS England, Seven Days a Week Forum, 2013c. Costing seven day services.

NICE, 2017. Who we are. About NICE. [WWW Document]. URL https://www.nice.org.uk/about/who-we-are (accessed 3.9.17).

NICE, 2014. Guide to the processes of technology appraisal [WWW Document]. URL https://www.nice.org.uk/process/pmg19/chapter/introduction (accessed 11.30.16).

NICE, 2013. Guide to the methods of technology appraisal 2013 The-reference-case [WWW Document]. URL https://www.nice.org.uk/article/pmg9/chapter/the-reference-case (accessed 3.31.16).

Norton, E.C., 1992. Incentive regulation of nursing homes. J. Health Econ. 11, 105–128.

Office for National Statistics, 2014. England, National Life Tables, 1980-82 to 2010-12.

Office for National Statistics, 2011a. England, Iterim Life Tables, 1980-82 to 2008-10 [Online].

Office for National Statistics, 2011b. Supporting Information: Lower Layer Super Output Area [WWW Document]. URL http://www.datadictionary.nhs.uk/data_dictionary/nhs_business_definitions/l/lower_layer_super_output_area_de.asp?shownav=1 (accessed 12.2.15).

Office for National Statistics, 2011c. Super Output Area (SOA) [WWW Document]. Off. Natl. Stat. URL http://www.ons.gov.uk/ons/guide-method/geography/beginner-s-guide/census/super-output-areas--soas-/index.html (accessed 12.2.15).

O'Neill, S., Kreif, N., Grieve, R., Sutton, M., Sekhon, J.S., 2016. Estimating causal effects: considering three alternatives to difference-in-differences estimation. Health Serv. Outcomes Res. Methodol. 16, 1–21. doi:10.1007/s10742-016-0146-8

Paris, V., Devaux, M., Wei, L., 2010. Health Systems Institutional Characteristics: A Survey of 29 OECD Countries (No. 50), OECD Health Working Papers.

Parke, D.W., 2007. Impact of a pay-for-performance intervention: financial analysis of a pilot program implementation and implications for ophthalmology. Trans. Am. Ophthalmol. Soc. 105, 448–460.

Population Data BC, 2016. Population Data BC [WWW Document]. URL https://www.popdata.bc.ca/home (accessed 6.8.15).

Quan, H., Sundararajan, V., Halfon, P., Fong, A., Burnand, B., Luthi, J.-C., Saunders, L.D., Beck, C.A., Feasby, T.E., Ghali, W.A., 2005. Coding algorithms for defining comorbidities in ICD-9-CM and ICD-10 administrative data. Med. Care 43, 1130–1139.

Raine, R., Fitzpatrick, R., Barratt, H., Bevan, G., Black, N., Boaden, R., Bower, P., Campbell, M., Denis, J.-L., Devers, K., Dixon-Woods, M., Fallowfield, L., Forder, J., Foy, R., Freemantle, N., Fulop, N.J., Gibbons, E., Gillies, C., Goulding, L., Grieve, R., Grimshaw, J., Howarth, E., Lilford, R.J., McDonald, R., Moore, G., Moore, L., Newhouse, R., O'Cathain, A., Or, Z., Papoutsi, C., Prady, S., Rycroft-Malone, J., Sekhon, J., Turner, S., Watson, S.I., Zwarenstein, M., 2016. Challenges, solutions and future directions in the evaluation of service innovations in health care and public health, Health Services and Delivery Research. NIHR Journals Library, Southampton (UK).

Reed Johnson, F., 2009. Editorial: Moving the QALY Forward or Just Stuck in Traffic? Value Health 12, S38–S39. doi:10.1111/j.1524-4733.2009.00521.x

Rosenthal, M.B., Frank, R.G., 2006. What is the empirical basis for paying for quality in health care? Med. Care Res. Rev. MCRR 63, 135–157. doi:10.1177/1077558705285291

Rosenthal, M.B., Li, Z., Robertson, A.D., Milstein, A., 2009. Impact of financial incentives for prenatal care on birth outcomes and spending. Health Serv. Res. 44, 1465–1479. doi:10.1111/j.1475-6773.2009.00996.x

Ryan, A.M., 2009. Effects of the Premier Hospital Quality Incentive Demonstration on Medicare Patient Mortality and Cost. Health Serv. Res. 44, 821–842. doi:10.1111/j.1475-6773.2009.00956.x

Ryan, A.M., Burgess, J.F., Dimick, J.B., 2015. Why We Should Not Be Indifferent to Specification Choices for Difference-in-Differences. Health Serv. Res. 50, 1211–1235. doi:10.1111/1475-6773.12270

Sanders, G.D., Neumann, P.J., Basu, A., Brock, D.W., Feeny, D., Krahn, M., Kuntz, K.M., Meltzer, D.O., Owens, D.K., Prosser, L.A., Salomon, J.A., Sculpher, M.J., Trikalinos, T.A., Russell, L.B., Siegel, J.E., Ganiats, T.G., 2016. Recommendations for Conduct, Methodological Practices, and Reporting of Cost-effectiveness Analyses: Second Panel on Cost-Effectiveness in Health and Medicine. JAMA 316, 1093–1103. doi:10.1001/jama.2016.12195

Schmulewitz, L., Proudfoot, A., Bell, D., 2005. The impact of weekends on outcome for emergency patients. Clin. Med. 5, 621–625. doi:10.7861/clinmedicine.5-6-621

Scott, A., Sivey, P., Ait Ouakrim, D., Willenberg, L., Naccarella, L., Furler, J., Young, D., 2011. The effect of financial incentives on the quality of health care provided by primary care physicians. Cochrane Database Syst. Rev. CD008451. doi:10.1002/14651858.CD008451.pub2

Shah, K., Praet, C., Devlin, N., Sussex, J., Appleby, J., Parkin, D., 2012. Is the aim of the English health care system to maximize QALYs? J. Health Serv. Res. Policy 17, 157–163. doi:10.1258/jhsrp.2012.011098

Shiell, A., Hawe, P., Gold, L., 2008. Complex interventions or complex systems? Implications for health economic evaluation. BMJ 336, 1281–1283. doi:10.1136/bmj.39569.510521.AD

Smith, M.D., Drummond, M., Brixner, D., 2009. Moving the QALY Forward: Rationale for Change. Value Health 12, S1–S4. doi:10.1111/j.1524-4733.2009.00514.x

Sutton, M., Elder, R., Guthrie, B., Watt, G., 2010. Record rewards: the effects of targeted quality incentives on the recording of risk factors by primary care providers. Health Econ. 19, 1–13. doi:10.1002/hec.1440

Sutton, M., Nikolova, S., Boaden, R., Lester, H., McDonald, R., Roland, M., 2012. Reduced Mortality with Hospital Pay for Performance in England. N. Engl. J. Med. 367, 1821–1828. doi:10.1056/NEJMsa1114951

Tunis, S.R., Stryer, D.B., Clancy, C.M., 2003. Practical Clinical Trials: Increasing the Value of Clinical Research for Decision Making in Clinical and Health Policy. JAMA 290, 1624–1632. doi:10.1001/jama.290.12.1624

Turner, S., Goulding, L., Denis, J.-L., McDonald, R., Fulop, N.J., 2016. Major system change: a management and organisational research perspective. NIHR Journals Library.

Ukoumunne, O.C., Gulliford, M.C., Chinn, S., Sterne, J.A.C., Burney, P.G.J., Donner, A., 1999. Evaluation of health interventions at area and organisation level. BMJ 319, 376–379. doi:10.1136/bmj.319.7206.376

Walker, S., Mason, A.R., Claxton, K., Cookson, R., Fenwick, E., Fleetcroft, R., Sculpher, M., 2010. Value for money and the Quality and Outcomes Framework in primary care in the UK NHS. Br. J. Gen. Pract. 60, e213–e220. doi:10.3399/bjgp10X501859

Weatherly, H., Drummond, M., Claxton, K., Cookson, R., Ferguson, B., Godfrey, C., Rice, N., Sculpher, M., Sowden, A., 2009. Methods for assessing the cost-effectiveness of public health interventions: key challenges and recommendations. Health Policy Amst. Neth. 93, 85–92. doi:10.1016/j.healthpol.2009.07.012

Yao, G.L., Novielli, N., Manaseki-Holland, S., Chen, Y.-F., Klink, M. van der, Barach, P., Chilton, P.J., Lilford, R.J., 2012. Evaluation of a predevelopment service delivery intervention: an application to improve clinical handovers. BMJ Qual. Saf. 21, i29–i38. doi:10.1136/bmjqs-2012-001210