Measuring the quality and safety of hospital care using specialty-specific indicators based on routinely collected administrative data: a feasibility study

William Lanyon Palmer MA MSc

Department of Primary Care and Public Health
School of Public Health
Faculty of Medicine
Imperial College London

A thesis submitted for the degree of Doctor of Philosophy, March 2014
Declaration of originality

I declare that this thesis is my own work and, where content is not my own, this is appropriately referenced.

William Palmer

Copyright declaration

The copyright of this thesis rests with the author and is made available under a Creative Commons Attribution Non-Commercial No Derivatives licence. Researchers are free to copy, distribute or transmit the thesis on the condition that they attribute it, that they do not use it for commercial purposes and that they do not alter, transform or build upon it. For any reuse or redistribution, researchers must make clear to others the licence terms of this work

William Palmer

Acknowledgements

I would like to thank the input of a range of experts who kindly inputted into this research, including Alex Bottle, Carmen Tsang, Charlie Davie, Sue Eve-Jones, and Susan Bewley. While I have been able to attribute the contributions of these experts in the relevant chapters, it is not possible – due to word count alone – to fully recognise the contributions of my two supervisors. Charles Vincent provided sage advice throughout, while Paul Aylin intrinsically knew when to balance pushing me on delivery goals and when I instead needed to talk about nu-folk and cycling apparel. To Karen Taylor: I both blame and thank you for suggesting a PhD (predominantly the latter). Whilst a privilege to do, six years of evening and weekend studying is inescapably draining and I’m indebted to my parents, friends and family for supporting me. Alas, too many of you deserve named recognition to list here.

The PhD was funded, in-part, by the National Audit Office (NAO), who were also my employers during the conduct of the PhD. The NAO played no part in the design, conduct, interpretation and presentation of the results. I am grateful for the support of the Brocher Foundation (www.brocher.ch) during my research fellowship in February and March 2013.

I used Hospital Episode Statistics (HES) dataset, which the Dr Foster Unit at Imperial has permission to hold under Section 251 (formerly Section 60) granted by the National Information Governance Board for Health and Social Care (NIGB, formerly the Patient Information Advisory Group). The unit have approval for using these data for research from the South East Research Ethics Committee.
Abstract

Using administrative data to measure the quality and safety of hospital care offers many opportunities. However, progress has been limited to few countries and predominantly to a small subset of broad measures, such as Hospital Standardised Mortality Rates. In this thesis, I investigate the potential advantages and feasibility – in terms of validity and applicability – of specialty-specific indicators.

In the first part of my PhD work, I examine the case for specialty-specific indicators. I also present potential applications which overcome some of the existing shortcomings of previous uses of indicators based on administrative data.

In the next stage of the project I focus on assessing feasibility by focusing on two specialties – stroke and obstetric care – conducting systematic reviews and consulting with experts to develop two indicator sets. As part of this, I identified the shortcomings in current use of indicators in these specialties. To investigate the limitations of these indicators, I applied the indicator definitions to English hospital administrative data (Hospital Episode Statistics, HES) and evaluated whether they can be used to discriminate between hospitals based on their performance and, importantly, to understand the effect of differences in coding practice.

The final aspect of the research was to investigate alternative applications for the indicators which can overcome some of the shortcomings highlighted in both the prior analyses and existing literature. In doing so, I raise serious, robust shortcomings on the quality and safety of weekend care.

**Novel aspects of this research**

<table>
<thead>
<tr>
<th>New indicator sets:</th>
<th>First collation of obstetric and stroke indicators based on administrative data</th>
</tr>
</thead>
<tbody>
<tr>
<td>New application:</td>
<td>First application of some of these indicators to English data</td>
</tr>
<tr>
<td></td>
<td>Applying indicator bundles to improve validity</td>
</tr>
<tr>
<td>New methods:</td>
<td>Present innovative methods to investigate bias from coding practice</td>
</tr>
<tr>
<td>New insights:</td>
<td>Reveal previously unknown issues on variation of care in England</td>
</tr>
</tbody>
</table>


Table of contents

Declaration of originality ........................................................................................................ 2
Copyright declaration .................................................................................................................. 2
Acknowledgements .................................................................................................................... 2
Abstract ..................................................................................................................................... 3
Table of contents ....................................................................................................................... 4
List of figures .............................................................................................................................. 11
List of tables ............................................................................................................................... 13
List of abbreviations .................................................................................................................. 16

Chapter 1. Quality, safety and their measurement .................................................................... 20
Overview .................................................................................................................................... 21
1.1. Introduction to quality and safety in healthcare ................................................................. 22
  1.1.1. Improving healthcare .................................................................................................... 22
  1.1.2. Patient safety ................................................................................................................ 22
  1.1.3. Quality in healthcare .................................................................................................... 23
  1.1.4. Continuum of quality and safety .................................................................................. 23
1.2. Causes, counts and costs .................................................................................................... 24
  1.2.1. Conceptual models ....................................................................................................... 24
  1.2.2. Prevalence and preventability .................................................................................... 24
  1.2.3. Cost of adverse events ............................................................................................... 25
1.3. Interest of quality and patient safety ................................................................................ 27
  1.3.1. Policy focus ................................................................................................................ 28
  1.3.2. Hospital focus .............................................................................................................. 28
  1.3.3. Patient and media focus ............................................................................................. 29
1.4. Need for indicators ............................................................................................................ 30
  1.4.1. Indicators .................................................................................................................... 30
  1.4.2. Demand for indicators ............................................................................................... 32
  1.4.3. Strengths and weaknesses of methods for identifying and analysing adverse events 34
  1.4.4. No consensus on a preferred technique ..................................................................... 36
1.5. Implications for scope ....................................................................................................... 36

Chapter 2. Specialty-specific indicators .................................................................................. 37
Overview .................................................................................................................................... 38
2.1. Use of administrative data ............................................................................................... 39
  2.1.1. Introduction to administrative data ............................................................................. 39
  2.1.2. Use of data .................................................................................................................. 40
  2.1.3. The development of indicators .................................................................................. 41
2.2. Advantages of using administrative data ........................................................................ 42
  2.2.1. Longitudinal ............................................................................................................... 43
  2.2.2. Ease and cost-effectiveness of use ............................................................................. 43
  2.2.3. Timely ....................................................................................................................... 44
  2.2.4. Coverage ................................................................................................................... 44
  2.2.5. Standardised definitions and rules ............................................................................ 45
  2.2.6. Basis for Pay for Performance schemes ................................................................. 45
  2.2.7. Independent reporting .............................................................................................. 46
2.3. Limitation of PSIs ............................................................................................................. 46
  2.3.1. Attribution of sequence and causation .................................................................... 46
  2.3.2. Judgements of responsibility and preventability ....................................................... 47
  2.3.3. Problems in coding details ....................................................................................... 47
  2.3.4. Case-mix adjustment ............................................................................................... 47
2.3.5. Accuracy of coding........................................................................................................48
2.3.6. Some have not been applied to ICD-10........................................................................48
2.3.7. Limited evidence on effectiveness................................................................................49
2.4. Review of validity of indicators.........................................................................................49
2.4.1. Defining validity of indicators......................................................................................49
2.4.2. Previous literature on validity.......................................................................................50
2.5. Potential applications of indicators..................................................................................51
2.5.1. Types of reporting.........................................................................................................51
2.5.2. Applying sets of indicators.........................................................................................52
2.5.3. Primarily organisational-wide indicators to date.........................................................53
2.6. Advantages of reporting specialty-specific indicators.....................................................54
2.6.1. Introduction to concept of specialty.............................................................................54
2.6.2. Introduction to specialty-specific indicators...............................................................55
2.6.3. Difference between specialties....................................................................................55
2.6.4. Nature of specialties.....................................................................................................57
2.6.5. Disadvantages of specialty-specific.............................................................................58
2.7. Research objectives..........................................................................................................59
2.7.1. Hypotheses................................................................................................................59
2.7.2. Thesis outline..............................................................................................................59

Chapter 3. Project design and methods .................................................................................60

Overview ...............................................................................................................................61
3.1. Introduction to study design.............................................................................................62
3.1.1. Introduction to feasibility.............................................................................................62
3.1.2. Addressing limitation in indicator use.........................................................................62
3.2. Analytical framework.......................................................................................................63
3.2.1. Introduction................................................................................................................63
3.2.2. Previous indicator validation.......................................................................................65
3.2.3. The analytical framework............................................................................................66
3.3. Specialty selection.............................................................................................................68
3.3.1. Selection framework.....................................................................................................68
3.3.2. Selected specialties......................................................................................................69
3.4. Developing an indicator set..............................................................................................70
3.4.1. Framework for developing indicators.........................................................................70
3.4.2. Literature review type................................................................................................72
3.4.3. Previous systematic reviews.......................................................................................73
3.4.4. Literature search resources........................................................................................74
3.5. Selecting indicators..........................................................................................................75
3.5.1. Considerations on indicator selection.........................................................................75
3.5.2. Defining indicators.......................................................................................................77
3.6. Applying indicators..........................................................................................................77
3.6.1. Descriptive statistics....................................................................................................78
3.6.2. Regression models.......................................................................................................79
3.6.3. Generalised linear models..........................................................................................79
3.6.4. Unadjusted indicator rates..........................................................................................80
3.6.5. Risk adjustment..........................................................................................................80
3.6.6. Goodness of fit............................................................................................................80
3.6.7. Standardisation of indicator results..........................................................................81
3.6.8. Funnel plots.................................................................................................................81
3.7. Setting...............................................................................................................................82
3.7.1. Structure of English NHS............................................................................................82
3.7.2. Quality and safety in England....................................................................................82
| 3.7.3.  | Focus on acute care ............................................................... | 83 |
| 3.8.    | Data ......................................................................................... | 84 |
| 3.8.1.  | Introduction to Hospital Episode Statistics ........................ | 84 |
| 3.8.2.  | Ethical considerations ............................................................. | 85 |
| 3.8.3.  | Software ..................................................................................... | 85 |

Chapter 4. Stroke literature review .................................................................. 87

Overview ........................................................................................................... 88

4.1. Background on stroke care ........................................................................ 89

4.1.1. About stroke ...................................................................................... 89

4.1.2. Importance ......................................................................................... 89

4.2. Introduction .............................................................................................. 91

4.2.1. Stroke as a specialty ......................................................................... 91

4.2.2. Stroke measurement ........................................................................ 91

4.3. Objectives ............................................................................................... 92

4.4. Methods .................................................................................................... 92

4.4.1. Search resources .............................................................................. 92

4.4.2. Validation of search results ............................................................. 92

4.4.3. Definitions ......................................................................................... 92

4.4.4. Search strategy ................................................................................. 93

4.5. Review process .......................................................................................... 94

4.5.1. First stage review criteria ................................................................ 94

4.5.2. Second stage review criteria ........................................................... 95

4.5.3. Data extraction and review ............................................................... 96

4.6. Results ...................................................................................................... 97

4.6.1. Identification of studies .................................................................... 97

4.6.2. Details from studies ......................................................................... 100

4.6.3. Study designs .................................................................................... 100

4.6.4. Details from datasets ....................................................................... 102

4.6.5. Risk models ....................................................................................... 104

4.6.7. Details from indicators .................................................................... 107

4.7. Discussion ............................................................................................... 109

4.7.1. Principle findings ............................................................................. 109

4.7.2. Source of data .................................................................................. 109

4.7.3. Stroke case ascertainment strategies .............................................. 109

4.7.4. Indicator validity ............................................................................... 110

4.7.5. Strengths and limitations ................................................................. 111

4.7.6. Conclusion ......................................................................................... 112

Chapter 5. Stroke variation and validation .......................................................... 115

Overview ........................................................................................................... 116

5.1. Background ............................................................................................ 117

5.1.1. Stroke care in England .................................................................... 117

5.1.2. Revisiting limitations in existing literature ...................................... 118

5.2. Objectives .............................................................................................. 118

5.3. Methods .................................................................................................. 118

5.3.1. Selecting indicators ......................................................................... 118

5.3.2. Developing indicator definitions ...................................................... 119

5.4. Application and validation of indicators ................................................ 120

5.4.1. Assumptions .................................................................................... 120

5.4.2. Risk-adjustment ............................................................................... 122

5.4.3. Statistical testing and validation ....................................................... 122

5.4.4. Sensitivity analysis for coding practice ............................................ 122
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Background on obstetrics</td>
<td>167</td>
</tr>
<tr>
<td>Importance</td>
<td>167</td>
</tr>
<tr>
<td>Obstetrics measurement</td>
<td>168</td>
</tr>
<tr>
<td>Objectives</td>
<td>169</td>
</tr>
<tr>
<td>Methods</td>
<td>169</td>
</tr>
<tr>
<td>8.3.1. Search resources and validation</td>
<td>169</td>
</tr>
<tr>
<td>8.3.2. Definitions</td>
<td>169</td>
</tr>
<tr>
<td>8.3.3. Search strategy</td>
<td>171</td>
</tr>
<tr>
<td>8.3.4. Search terms</td>
<td>171</td>
</tr>
<tr>
<td>8.3.5. Review process</td>
<td>173</td>
</tr>
<tr>
<td>Results</td>
<td>175</td>
</tr>
<tr>
<td>8.4.1. Identification of studies that applied administrative data</td>
<td>177</td>
</tr>
<tr>
<td>8.4.2. Details from studies applying administrative data</td>
<td>179</td>
</tr>
<tr>
<td>8.4.3. Details from datasets</td>
<td>181</td>
</tr>
<tr>
<td>8.4.4. Risk models</td>
<td>182</td>
</tr>
<tr>
<td>8.4.5. Details from indicators</td>
<td>182</td>
</tr>
<tr>
<td>Discussion</td>
<td>184</td>
</tr>
<tr>
<td>8.5.1. Principle findings</td>
<td>184</td>
</tr>
<tr>
<td>8.5.2. Indicator validity</td>
<td>185</td>
</tr>
<tr>
<td>8.5.3. Strengths and limitations</td>
<td>187</td>
</tr>
<tr>
<td>8.5.4. Conclusions</td>
<td>189</td>
</tr>
<tr>
<td>Chapter 9. Obstetrics variation and validation</td>
<td>191</td>
</tr>
<tr>
<td>Overview</td>
<td>192</td>
</tr>
<tr>
<td>9.1. Background</td>
<td>193</td>
</tr>
<tr>
<td>9.1.1. Obstetrics care in England</td>
<td>193</td>
</tr>
<tr>
<td>9.1.2. Revisiting existing literature</td>
<td>193</td>
</tr>
<tr>
<td>9.2. Objectives</td>
<td>196</td>
</tr>
<tr>
<td>9.3. Methods</td>
<td>196</td>
</tr>
<tr>
<td>9.3.1. Data</td>
<td>196</td>
</tr>
<tr>
<td>9.3.2. Selection of indicators</td>
<td>197</td>
</tr>
<tr>
<td>9.3.3. Developing indicator definitions</td>
<td>198</td>
</tr>
<tr>
<td>9.4. Application and validation of indicators</td>
<td>200</td>
</tr>
<tr>
<td>9.4.1. Assumptions</td>
<td>200</td>
</tr>
<tr>
<td>9.4.2. Risk adjustment</td>
<td>202</td>
</tr>
<tr>
<td>9.4.3. Validation</td>
<td>202</td>
</tr>
<tr>
<td>9.4.4. Amenable: associations between organisational factors and performance</td>
<td>203</td>
</tr>
<tr>
<td>9.4.5. Alignment with analytical framework</td>
<td>204</td>
</tr>
<tr>
<td>9.4.6. Analytical packages</td>
<td>205</td>
</tr>
<tr>
<td>9.5. Results</td>
<td>205</td>
</tr>
<tr>
<td>9.5.1. Importance</td>
<td>205</td>
</tr>
<tr>
<td>9.5.2. Amenable</td>
<td>207</td>
</tr>
<tr>
<td>9.5.3. Precision</td>
<td>208</td>
</tr>
<tr>
<td>9.5.4. Construct validity</td>
<td>212</td>
</tr>
<tr>
<td>9.5.5. Minimum bias: coding practice</td>
<td>213</td>
</tr>
<tr>
<td>9.6. Discussion</td>
<td>216</td>
</tr>
<tr>
<td>9.6.1. Summary of findings</td>
<td>216</td>
</tr>
<tr>
<td>9.6.2. Amenable: scope for improvement</td>
<td>216</td>
</tr>
<tr>
<td>9.6.3. Precision</td>
<td>216</td>
</tr>
<tr>
<td>9.6.4. Minimum bias</td>
<td>217</td>
</tr>
<tr>
<td>9.6.5. Errors in underlying data</td>
<td>218</td>
</tr>
<tr>
<td>9.6.6. Face and content validity</td>
<td>219</td>
</tr>
</tbody>
</table>
Chapter 10. Obstetrics temporal trends

Overview

10.1. Background

10.1.1. Existing obstetric studies on administrative data

10.1.2. Other literature

10.2. Objectives

10.3. Methods

10.3.1. General methodological considerations

10.3.2. Additional ‘day of the week’ methodological considerations

10.4. Results

10.4.1. Longitudinal analysis

10.4.2. Association with day of the week

10.5. Discussion

10.5.1. Summary of findings on seasonality

10.5.2. Day of week

10.5.3. Conclusion

Chapter 11. Discussion

11.1. Summary

11.1.1. Overview

11.1.2. Summary of findings, by hypothesis

11.2. Methodological issues

11.2.1. Limitations to development of indicator sets

11.2.2. Indicator validation

11.2.3. Limitations to findings on shortcomings in care

11.3. Future research

11.3.1. Developing indicator sets for other specialties

11.3.2. Further research within these specialties

11.3.3. Developing reporting standards

11.3.4. Study designs

11.3.5. Introduction

11.3.5.1. Epidemiological studies

11.3.5.2. Quasi-experimental

11.4. Data implications

11.4.1. Background

11.4.2. Recommended interventions

11.5. Clinical implications

11.5.1. Clinical guidelines and quality improvement

11.5.2. Advantages of reporting to a ‘specialty team’

11.5.2.2. Learning and culture

11.5.2.3. Develop new and existing indicators

11.6. Policy implications

11.6.1. Variations in performance

11.6.1.2. Use of benchmarking

11.6.2. Addressing issue on association between complications and day of the week

11.6.3. Application of indicators by policy makers

11.6.3.3. Types of application
11.6.3.4. Accountability ......................................................... 267
11.6.3.5. Reimbursement ......................................................... 267
11.6.3.6. Make judgements about priorities ................................ 268
11.6.3.7. Case findings .......................................................... 268
11.7. Conclusion ....................................................................... 269

References .............................................................................. 270

Appendices ............................................................................... 296
  Appendix A. Outputs from the project ........................................ 296
  Appendix B. Details of previous literature reviews of quality and safety indicators ........................................ 297
  Appendix C: Additional websites searched .................................. 301
  Appendix D: Stroke indicators .................................................... 304
  Appendix E: Obstetrics indicators ............................................... 306
List of figures

Figure 1. Trend in publications based on administrative datasets ................................................................. 40
Figure 2. Trends in publication of quality indicators and use of administrative data ........................................ 41
Figure 3. Dimensions of validity used by AHRQ .................................................................................................. 50
Figure 4. Suggested indicator applications ......................................................................................................... 52
Figure 5. Indicative recommended stroke care pathway ................................................................................... 90
Figure 6. Search terms .......................................................................................................................................... 94
Figure 7. Inclusion and exclusion of articles in the literature review ................................................................. 98
Figure 8. Number of publications compared to latest year of data ................................................................... 100
Figure 9. Positioning of the stroke indicators across the care pathway ............................................................. 119
Figure 10: Details of stroke indicators used in analysis ...................................................................................... 120
Figure 11: Assumptions for identifying stroke using Hospital Episode Statistics ......................................... 121
Figure 12: Funnel plots of hospital-level performance across six indicators spanning the hospital stroke care pathway for 2009-10 ................................................................................................. 126
Figure 13. Performance by month¹ .................................................................................................................. 142
Figure 14. Example of hospital trends ................................................................................................................ 147
Figure 15. Number of stroke admissions, by day of presentation ................................................................. 153
Figure 16. Performance by day of admission ..................................................................................................... 157
Figure 17. Hospital-level standardised mortality ratio, for deaths within 7 days of admission on a weekend in 2009/10 ......................................................................................................................... 163
Figure 18. Overview of the obstetric care pathway ........................................................................................... 167
Figure 19. Search strategy ................................................................................................................................. 173
Figure 20. Inclusion and exclusion of articles in the literature review ........................................................... 176
Figure 21. Characteristics of literature ............................................................................................................... 177
Figure 22. Trend in annual publication numbers .............................................................................................. 178
Figure 23. Studies comparing hospital performance using obstetric QIs ......................................................... 195
Figure 24. Obstetric indicators across the care pathway .................................................................................... 198
Figure 25. Key diagnosis fields and codes for identifying delivery episode .................................................. 200
Figure 26. Coding rule examples ...................................................................................................................... 201
Figure 27: Funnel plots of trust-level performance across the obstetric indicators, 2011-12 .................. 209
Figure 28: Funnel plots of site-level performance across the obstetric indicators, 2011-12 ............... 210
Figure 29: Trends in obstetric indicators, by month ....................................................................................... 232
Figure 30. Proportion of births and maternities and breakdown in number of elective caesarean sections, by day of week ........................................................................................................................................................................ 234
Figure 31: Association between performance and day of admission/birth .................................................. 237
Figure 32: Potential uses of patient safety indicators .................................................................................. 265
List of tables

Table 1: National estimates on the prevalence of patient safety events .................................................. 25
Table 2: Examples of different indicators .................................................................................................. 31
Table 3: The strengths and weaknesses of methods for measuring quality and safety .............................. 35
Table 4. Examples of key quality and patient safety indicator sets ......................................................... 42
Table 5. Methods table .......................................................................................................................... 59
Table 6. Recommended study design to mitigate limitations ................................................................... 63
Table 7. Domains of validity .................................................................................................................. 65
Table 8. Analytical framework ................................................................................................................ 67
Table 9. Differences in selected specialties ............................................................................................ 69
Table 10. Development of indicator sets ................................................................................................ 71
Table 11. Key distinctions between narrative and systematic reviews .................................................. 73
Table 12. Selected details from sample of previous review of indicators ................................................. 74
Table 13. Use of multiple measures and mix of process and outcome measures to address analytical framework .................................................................................................................. 77
Table 14. Statistical tests to address the analytical framework ............................................................... 78
Table 15. Simple statistical tests ............................................................................................................. 78
Table 16. Information extracted .............................................................................................................. 96
Table 17. Characteristics of literature .................................................................................................... 99
Table 18. Variations in use of ICD-10 diagnosis codes for identifying stroke ........................................ 103
Table 19. Factors included in risk models ............................................................................................... 104
Table 20. Summary of key factors used in studies ................................................................................ 106
Table 21. Categories of measures .......................................................................................................... 108
Table 22. Summary of implications for study design .............................................................................. 114
Table 23. Analytical framework ............................................................................................................. 124
Table 24: Number of trusts identified as having statistically significant above or below average performance .......................................................................................................................... 127
Table 25: Coefficient of correlation between pairs of indicators ............................................................ 128
Table 26: Model fit statistics ................................................................................................................... 129
Table 27: Annual trend in performance ................................................................................................... 141
Table 28. Summary of differences in performance ................................................................................. 143
Table 29. Crude rates before and after intervention ............................................................................... 144
Table 30. Time-Series regression analysis of stroke indicators in London relative to the rest of England................................................................. 145
Table 31. Characteristics of the study population ......................................................................................... 153
Table 32. Association between weekday/weekend admission and indicators of quality and safety of care ................................................................................................................................. 155
Table 33. The variation in ‘weekend effect’ across different age groups, for 7-day in-hospital mortality .............................................................................................................................................. 158
Table 34. The specialty of the first responsible consultant after onset of stroke ........................................... 159
Table 35. Search terms ............................................................................................................................................... 171
Table 36. Information extracted ................................................................................................................................. 174
Table 37. Example papers ........................................................................................................................................... 180
Table 38. Key obstetrics indicators ............................................................................................................................... 182
Table 39. Risk adjustment in the SimPatIE obstetrics measures ........................................................................... 186
Table 40. Selected benchmarks for performance ........................................................................................................ 187
Table 41. Summary of implications of study design .............................................................................................. 190
Table 42. Details of obstetric indicators used in analysis ......................................................................................... 198
Table 43. Analytical framework ................................................................................................................................. 205
Table 44. Characteristics of mothers in the study population .................................................................................... 206
Table 45: Association between birth-to-midwife ratio and performance measures .................................................. 208
Table 46: Number of trusts identified as having statistically significant above or below average performance ............................................................................................................................................... 211
Table 47: Trust-level Coefficient of correlation between pairs of indicators ....................................................... 212
Table 48: Site-level coefficient of correlation between pairs of indicators ........................................................... 213
Table 49: Correlation between adjusted and unadjusted rates, and model fit statistics .............................................. 214
Table 50: Effect of including coding practice in case-mix model ................................................................................. 215
Table 51. Comparison of numbers of maternities and births from the HES extract and ONS births register ............................................................................................................................................... 219
Table 52. Studies using obstetric indicators to investigate temporal issues .......................................................... 227
Table 53. Obstetric indicators ....................................................................................................................................... 228
Table 54. Recommended levels of consultant presence .......................................................................................... 230
Table 55. Number of births and maternities and complication rates ........................................................................ 231
Table 56. Longitudinal trend and seasonal effect \(^1\) .............................................................................................. 233
Table 57. Characteristics of the study population ....................................................................................................... 234
Table 58. Association between weekday/weekend admission (or birth) and indicators of quality and safety of care.................................................................236
Table 59 Perinatal rates by delivery rates for common weekend delivery methods.................. 238
Table 60. Association between business and complication rates, for weekdays April 2010 – March 2012 ........................................................................................................239
Table 61. Association between compliance with consultant staffing levels and indicators of quality and safety of care.................................................................240
Table 62. Systematic review findings in relation to analytical framework.......................... 253
Table 63. Summary of strengths and weaknesses of the hospital and patient-level approaches ..... 257
Table 64. Key strengths and weakness of common applications of administrative data..............266
<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Full Form</th>
</tr>
</thead>
<tbody>
<tr>
<td>Agency for Healthcare Research and Quality</td>
<td>AHRQ</td>
</tr>
<tr>
<td>Agency for Healthcare Research and Quality Patient Safety Indicators</td>
<td>AHRQ PSIs</td>
</tr>
<tr>
<td>Australian Council for Safety and Quality in Health Care</td>
<td>ACSQ</td>
</tr>
<tr>
<td>Caesarean Section</td>
<td>CS</td>
</tr>
<tr>
<td>Canadian Institute for Health Information</td>
<td>CIHI</td>
</tr>
<tr>
<td>Care Quality Commission</td>
<td>CQC</td>
</tr>
<tr>
<td>Centre for Patient Safety and Service Quality</td>
<td>CPSSQ</td>
</tr>
<tr>
<td>Cochrane Library: Database of Systematic Reviews</td>
<td>CDSR</td>
</tr>
<tr>
<td>Complication Screening Program</td>
<td>CSP</td>
</tr>
<tr>
<td>Danish National Indicator Project</td>
<td>NIP</td>
</tr>
<tr>
<td>Database of Abstracts of Reviews of Effectiveness</td>
<td>DARE</td>
</tr>
<tr>
<td>Deep vein thrombosis</td>
<td>DVT</td>
</tr>
<tr>
<td>Department of Health (England)</td>
<td>DH</td>
</tr>
<tr>
<td>Do Not Resuscitate</td>
<td>DNR</td>
</tr>
<tr>
<td>Excerpta Medica Database</td>
<td>EMBASE</td>
</tr>
<tr>
<td>Gross Domestic Product</td>
<td>GDP</td>
</tr>
<tr>
<td>Health Care Financing Administration</td>
<td>HCFA</td>
</tr>
<tr>
<td>Health Maintenance Organisation</td>
<td>HMO</td>
</tr>
<tr>
<td>Health Technology Assessment</td>
<td>HTA</td>
</tr>
<tr>
<td>Healthcare Cost and Utilization Project</td>
<td>HCUP</td>
</tr>
<tr>
<td>Healthcare Cost and Utilization Project National Inpatient Sample</td>
<td>HCUP NIS</td>
</tr>
<tr>
<td>Healthcare Cost and Utilization Project Quality Indicators</td>
<td>HCUP QIs</td>
</tr>
<tr>
<td>Healthcare Cost and Utilization Project Refined Quality Indicators</td>
<td>HCUP II QIs</td>
</tr>
<tr>
<td>Healthcare Cost and Utilization Project State Inpatient Databases</td>
<td>HCUP SIDs</td>
</tr>
<tr>
<td>Hospital Episode Statistics</td>
<td>HES</td>
</tr>
<tr>
<td>Hospital Standardised Mortality Rates</td>
<td>HSMR</td>
</tr>
<tr>
<td>Institute for Healthcare Improvement</td>
<td>IHI</td>
</tr>
<tr>
<td>Institute of Medicine</td>
<td>IOM</td>
</tr>
<tr>
<td>Intensive Care Unit</td>
<td>ICU</td>
</tr>
<tr>
<td>International Statistical Classification of Diseases and Related Health Problems</td>
<td>ICD</td>
</tr>
<tr>
<td>International Classification of Diseases, Ninth Revision, Clinical Modification</td>
<td>ICD-9-CM</td>
</tr>
<tr>
<td>International Classification of Diseases, Tenth Revision</td>
<td>ICD-10</td>
</tr>
<tr>
<td>International Quality Indicator Project</td>
<td>IQIP Project</td>
</tr>
<tr>
<td>Joint Commission on Accreditation of Healthcare Organizations</td>
<td>JCAHO</td>
</tr>
<tr>
<td>Length of stay</td>
<td>LOS</td>
</tr>
<tr>
<td>Medical Literature Analysis and Retrieval System Online</td>
<td>MEDLINE</td>
</tr>
<tr>
<td>Medicare Provider Analysis and Review file</td>
<td>MEDPAR</td>
</tr>
<tr>
<td>National Health Service</td>
<td>NHS</td>
</tr>
<tr>
<td>National Institute for Health Research</td>
<td>NIHR</td>
</tr>
<tr>
<td>National Institute for Health and Clinical Excellence</td>
<td>NICE</td>
</tr>
<tr>
<td>Danish National Indicators Project</td>
<td>NIP</td>
</tr>
<tr>
<td>National Patient Safety Agency</td>
<td>NPSA</td>
</tr>
<tr>
<td>National Reporting and Learning System</td>
<td>NRLS</td>
</tr>
<tr>
<td>Organization for Economic Co-operation and Development</td>
<td>OECD</td>
</tr>
<tr>
<td>OECD Health Care Quality Indicator Project</td>
<td>OECD HCQ</td>
</tr>
<tr>
<td>Patient safety indicator(s)</td>
<td>PSI(s)</td>
</tr>
<tr>
<td>Postpartum haemorrhage</td>
<td>PPH</td>
</tr>
<tr>
<td>Term</td>
<td>Abbreviation</td>
</tr>
<tr>
<td>-------------------------------------------</td>
<td>--------------</td>
</tr>
<tr>
<td>Present on admission</td>
<td>POA</td>
</tr>
<tr>
<td>Quality Adjusted Life Years</td>
<td>QALYs</td>
</tr>
<tr>
<td>Quality Indicators</td>
<td>QIs</td>
</tr>
<tr>
<td>Safety Improvement for Patients In Europe Project</td>
<td>SimPatIE Project</td>
</tr>
<tr>
<td>Transient Ischaemic Attack</td>
<td>TIA</td>
</tr>
<tr>
<td>United Kingdom</td>
<td>UK</td>
</tr>
<tr>
<td>United States America</td>
<td>USA</td>
</tr>
<tr>
<td>Urinary Tract Infection</td>
<td>UTI</td>
</tr>
<tr>
<td>Vaginal Birth</td>
<td>VB</td>
</tr>
<tr>
<td>Vaginal birth with instrument</td>
<td>Vbwi</td>
</tr>
<tr>
<td>Vaginal birth without instrument</td>
<td>Vbwol</td>
</tr>
<tr>
<td>Veterans Association</td>
<td>VA</td>
</tr>
<tr>
<td>World Health Organization</td>
<td>WHO</td>
</tr>
</tbody>
</table>
Jumbo (1861 – 1885) was an African elephant of enormous proportions. The giant mammal was famed in Paris and London zoos. Despite 100,000 school children writing to Queen Victoria not to sell him, Jumbo was bought by the owner of “The Greatest Show on Earth”. The picture has been included here to introduce an idiom, *the elephant in the room*, and a fable, *the blind men and the elephant*, from which you can draw analogies to measuring healthcare quality and safety. However, this thesis perhaps offers a more promising story than that of Jumbo. Unfortunately, the 4 metre tall Jumbo died – reportedly saving a young circus elephant – when crushed by a locomotive. Further, the stuffed Jumbo was destroyed in a fire in 1975.
Part One:

Introduction and concepts

*It was six men of Indostan*

*To learning much inclined,*

*Who went to see the Elephant*

*(Though all of them were blind),*

*That each by observation*

*Might satisfy his mind.*

(The Blind Men and the Elephant, John Godfrey Saxe 1816-87)
Chapter 1.

Quality, safety and their measurement
Overview

"Declare the past, diagnose the present, foretell the future; practice these acts.

As to diseases, make a habit of two things-to help, or at least to do no harm"

(Epidemics, Book 1, Section XI, Hippocrates, 460-370BC)

Context

This chapter sets out the quality and safety landscape outlining the impact of, and interest in, shortcomings in healthcare. In particular, I highlight the importance of measuring the quality and safety of hospital care. Thereafter, I present the previous techniques used to meet this measurement demand, and provide a short critique of these. The chapter closes by setting out the case for further development of measures of quality and safety.

Methods

A narrative review of the literature

Findings

The quality and safety of healthcare remains a global concern. The literature reveals: significant shortcomings in quality and safety; high impact of these adverse events; potential preventability; and widespread interest. There is scope to develop further healthcare indicators and apply them in a broader array of countries than currently adopts them.

Related papers

1.1. Introduction to quality and safety in healthcare

1.1.1. Improving healthcare

There has been considerable progress in health care across the world. From 2000 to 2009, total expenditure on health has increased by a seventh (from 8.2% to 9.4% of total global GDP) and now represents a more significant part of governments accounts, with general government expenditure on health increasing from 13.3% to 14.3% of total government expenditure.¹ People’s health is also improving. The World Health Organisation (WHO) estimated that the reduction in the rate of child mortality between 1978 and 2006 is equivalent to over 18 000 children’s lives being saved every day.² Global life expectancy has also increased by four years, to 68, between 1990 and 2009.¹

1.1.2. Patient safety

A seminal report by the Institute of Medicine (IoM), published in 1999, outlined the patient safety issue.³ Despite these improvements, there is a longstanding acknowledgment of the existence of persistent risks to patients’ safety. ‘The longstanding cornerstone of medicine “do no harm” exists because of the fragility of life and health during medical care encounters’⁴ While health services are treating more patients, due to the ever increasing size and complexity of healthcare, upholding this maxim is proving more and more difficult.⁵

Such harm manifests as patient safety events, defined by one organisation as “any unintended or unexpected incident which could have or did lead to harm for one or more patients receiving... care”.⁶ Patient safety events come under many labels, including: iatrogenic harm, medical errors, nosocomial infections, substandard care, adverse events, critical incidents, preventable or avoidable complications, and sentinel events.

The number of labels reflects the various types of adverse events – such as diagnostic errors, equipment failure, infections, blood transfusion-related injuries, and misinterpretation of other medical orders ⁷ – and belies the common belief that medical errors usually involve incorrect drug use or wrong-site surgery. Whilst the context within which healthcare services are operating is making the risks more pronounced, the problem of patient safety events is not new, with studies on such incidents being produced throughout the 1950s and an emerging body of evidence by the 1990s.⁸ However, despite these early reports, in terms of research, patient safety is still a new and fast growing area.⁹
1.3. Quality in healthcare

As a result of developments of the health services and increased scrutiny, expectations have changed. Healthcare is expected to be not just safe but of high quality which can be defined as “the degree to which health services for individuals and populations increase the likelihood of desired health outcomes and are consistent with current professional knowledge.” This can include new technologies and procedures. The proliferation of health research and resultant guidelines to instruct the provision of care has helped to codify differences in quality, in terms of processes as well as outcomes. Again ‘quality’ is a broad term and can include not only evaluation of clinical contact but of patient satisfaction and access.

1.4. Continuum of quality and safety

1.4.1. Difficulty in disaggregating quality and safety

The changes in perceptions over what is expected from healthcare services redefine the line between quality and patient safety. To demonstrate this problem, consider the thought experiment on misdiagnosis:

Patient A has a rare condition that has ambiguous symptoms. Patient B has a typical case of a common disease that has high diagnosis rates within the population. Both conditions are easily treatable but, on presentation to the health service, both are misdiagnosed and the prescribed treatment makes no effect on Patient A’s state but makes Patient B less well.

Given Patient B might have reasonable expectations of being correctly diagnosed and the failure to do so caused harm, most definitions would categorise this as a patient safety event. Whereas, the lower expectations of correct diagnosis and lack of harm would, under most definitions, classify Patient A as quality, but not patient safety, incident. However, the question of what likelihood of correct diagnosis and level of harm (or failure to improve) does the line between quality and safety fall remains subjective. Indeed, this lack of distinction has been highlighted in academic literature.\textsuperscript{11,12}

1.4.2. Defining quality and safety in this project

Despite difficulties in disaggregating between these two aspects of healthcare delivery, some parts of the remainder of this chapter (particular on estimating prevalence and costs) will focus predominantly on commonly evoked patient safety events and measurement since this is sufficient to describe the importance of measurement, and benefits from being easier to conceptualise given the more dichotomous nature of patient safety events. I did not
attempt to create a potentially burdensome and redundant framework for codifying events along the continuum since both quality and safety have intrinsic interest to all stakeholders, as detailed in section 1.3, p27. For brevity, I use the term ‘adverse event’ to cover both quality and patient safety incidents.

1.2. Causes, counts and costs

1.2.1. Conceptual models
Many frameworks have been used to describe the causes of patient safety events or shortcomings in quality, although most have focused on the former, such as: Donabedian’s interconnected outcome, process and structure;\textsuperscript{13} the widely-used\textsuperscript{14} categorisation proposed by James Reason with errors as slips, lapse or mistakes;\textsuperscript{15} or the long list of causes presented by the World Alliance for Patient Safety including poor training, misdiagnosis and latent failures.\textsuperscript{16} Given the different dimensions that can compound to result in an adverse event, any categorisation will struggle to be both mutually exclusive and comprehensively exhaustive. There is, therefore, no agreement on the best scheme for classifying adverse events\textsuperscript{14} and many existing categorisations have omissions. For example, Reason’s popular scheme fails to capture some commonly cited adverse events, such as equipment failures.

1.2.2. Prevalence and preventability

1.2.2.1. National estimates
The Harvard study in 1984 was the first major investigation into the extent of the patient safety issue and estimated that adverse events, defined in this instance as “injuries caused by medical management, and of the subgroup of such injuries that resulted from negligent or substandard care” (p370), occurred in 3.7% of admissions.\textsuperscript{17} A similar, subsequent study undertaken in Australia, using a broader definition of adverse events, estimated an incidence rate of 16.6% of admissions.\textsuperscript{18} Although these two studies represent the most exhaustive evaluation of the number of patient safety events in their respective countries, they are still thought to underestimate the actual prevalence.\textsuperscript{19} A number of further studies have estimated the prevalence of such events, as outlined in Table 1 below.
### Table 1: National estimates on the prevalence of patient safety events

<table>
<thead>
<tr>
<th>Country</th>
<th>Study</th>
<th>Number of hospital admissions †</th>
<th>Number of adverse events †</th>
<th>Adverse event rate (% patients) †</th>
<th>Adverse event rate (% admissions)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Australia</td>
<td>Quality in Australian Health Care Study, QAHCS (1992)</td>
<td>14 179</td>
<td>2 353</td>
<td>-</td>
<td>16.6</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(1 499) ‡</td>
<td></td>
<td>(10.6) ‡</td>
</tr>
<tr>
<td>France</td>
<td>(Pilot study)</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>14.5</td>
</tr>
<tr>
<td>New Zealand^</td>
<td>(1998)</td>
<td>6 579</td>
<td>849</td>
<td>12.9</td>
<td>11.2</td>
</tr>
<tr>
<td>United Kingdom</td>
<td>(1999-2000)</td>
<td>1 014</td>
<td>119</td>
<td>11.7</td>
<td>10.8</td>
</tr>
<tr>
<td>Denmark</td>
<td>(1998)</td>
<td>1 097</td>
<td>176</td>
<td>9.0</td>
<td>9.0</td>
</tr>
<tr>
<td>Canada</td>
<td>(2001)</td>
<td>3 720</td>
<td>279</td>
<td>7.5</td>
<td>7.5</td>
</tr>
<tr>
<td>United States</td>
<td>California Insurance Feasibility Study</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>4.7</td>
</tr>
<tr>
<td>United States</td>
<td>Harvard Medical Practice Study (1984)</td>
<td>30 195</td>
<td>1 133</td>
<td>3.8</td>
<td>3.7</td>
</tr>
<tr>
<td>United States</td>
<td>Utah-Colorado Study, UTCOS (1992)</td>
<td>14 565</td>
<td>475</td>
<td>3.2</td>
<td>2.9</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(787) ‡</td>
<td></td>
<td>(5.4) ‡</td>
</tr>
</tbody>
</table>

Notes: ^ Included all acute care; † Additional information from WHO (2004); ‡ UTCOS and QAHCS were revised using the same methodology. These comparable results are given in brackets; ~ Study measured ‘potentially compensable events’.

Source: Adapted from Vincent (2006), p42.

#### 1.2.2.2. Amenable to healthcare

Further analysis has suggested that most patient safety events identified in the studies of national prevalence were preventable, although “some adverse events are not preventable and they reflect the risk associated with treatment, such as a life-threatening allergic reaction to a drug when the patient had no known allergies to it.” In the Harvard case, later analysis estimated that 70% of adverse events were potentially preventable and in the Australian study the equivalent figure was 51%. In a preliminary study in 2001 using a similar method in England, of the adverse events identified (which occurred in 10.8% of admissions) almost half (48%) were judged preventable.

#### 1.2.3. Cost of adverse events

##### 1.2.3.1. Increased use of hospital resources

Patients who are harmed as a result of an adverse event can, and often do, have longer admissions and require additional procedures. For example, in the Harvard and Utah studies,
adverse drug events were associated with an increase of 2.2 and 1.9 days, respectively, in comparison to data from matched controls. When considering only preventable events, the figure for the Harvard study was 4.6 days. Analysis focusing on paediatric care showed that some patient safety events were associated with considerable increases in average length of stay, with an increase of 18 days for decubitus ulcers, 26 days for postoperative sepsis and 30 days for infection as a result of medical care. In addition to the acute care costs, there are also potentially other costs incurred as a result of adverse events, such as in primary care and attendances at accident and emergency when the harm, as can be the case for hospital acquired infection, presents itself after discharge.

1.2.3.2. Increased mortality
As well as harm to patients, adverse events can result in death. Analysing the effects on mortality over a range of patient safety events, Miller and Zhan (2004) found an increased rate of in-hospital deaths. The varying effects identified by the study ranged from those for the patient safety event of birth trauma, with an increased rate of mortality with odds ratio 1.3 compared with infants who did not have trauma at birth, to iatrogenic pneumothorax (odds ratio of 7.5) and postoperative respiratory failure (odds ratio of 76.6). The authors of this analysis, as with that presented in the previous paragraph, are not able to control for potential bias, such as in the scenario where an unmeasured case-mix factor is associated with an increased likelihood both of an adverse event being identified (due to more obvious poorer outcome in those with complex comorbidities) and death. In another study investigating the proportion of active-care patient deaths that were preventable, almost a quarter (23%) were rated by a clinician panel as at least possibly preventable by optimal care, with 6% deemed to be probably or definitely preventable.

1.2.3.3. Financial costs
Adverse events are also an important issue due to their financial cost. An event may introduce additional costs through, for example: the hospital being reimbursed for additional bed days and procedures; legal fees; a burden on the workforce to support the additional care required; emotional cost to any practitioner involved; and any physical or mental harm on the patient. To date, few studies have looked at the societal and personal costs with focus remaining on hospital charges. Drawing again on the findings of the Harvard and Utah studies, adverse drug events have been associated with an increased average cost per admission of $2 262 and $2 595 ($4 685 when considering only preventable adverse drug events). These figures are dwarfed by a US study on paediatric care which found,
across a range of patient safety events, excess charges per discharge ranged from $30 000 to $140 000 depending on the type of patient safety events.  

### 1.2.3.4. National estimates of costs

Further to estimating the effects on patient safety events at patient and hospital levels, efforts have been made to extrapolate the burden on a national scale. Many of these estimates are based on the studies, outlined in Table 1 (p25), that evaluated the prevalence of adverse events. An IoM report estimated that between 44 000 and 98 000 deaths per year were due to medical errors in US hospitals, whereas a news report stated that medical mistakes killed 180 000 people a year. However, McDonald and colleagues (2000) argued that these estimates are likely to be overestimates; in particular, since there is a lack of consideration of the expected risk of in-hospital death in the absence of a medical error. Further questions are cast over the validity of these estimates by the fact that the extrapolation are made by people not involved in the original reports and so the original notions of preventability are missed. Specifically, there are problems associated with the extrapolation of results from small studies to a national level and comparisons between countries are affected by the variations in definitions used in the original studies. A more sophisticated approach, which is relevant to the discussion at the end of the thesis (paragraph 11.6.3.6, p268), might be to assess quality-adjusted life years (QALYs) lost.

As well as the estimates based on the effects on mortality rates, a number of efforts have been made to estimate the cost of patient safety events on a national level. In the US, solely within paediatric care, excess charges were estimated to be in the region of $1 billion. One report suggests that the national annual cost of patient safety events is approximately $37.6 billion. Of this about half of the costs ($17 billion) are associated with preventable errors, and half of this expenditure (on preventable medical errors) is accounted for through direct health care costs. In Australia, a study based on 2003-04 data estimated that the burden was in the region of Aus$2 billion (equivalent to around US $1.5 billion).

### 1.3. Interest of quality and patient safety

There has been an increased focus on quality and safety from all stakeholders. This increased interest is a results of a multitude of factors, including: changing risks; dramatic reforms in healthcare systems and organisational structures; the introduction of legislation governing quality and safety and policies such as linking adverse events to healthcare
providers’ remuneration;\(^{33}\) and increased awareness - both in terms of certain types of events such as hospital acquired infections\(^{35}\) and due to high-profile adverse events.\(^{36}\)

### 1.3.1. Policy focus

#### 1.3.1.1. Government priority

Perhaps as a result of these startling figures on the national effects of poor quality and patient safety events, many governments have begun to increase the priority they place on reducing such events.\(^{37}\) Some governments have demonstrated their intentions by introducing incentives to reduce adverse events. One indication of this change is the number of new agencies and legislation designed to improve patient safety, such as the National Patient Safety Agency (NPSA) which was established in England in 2001 (although the function transferred to the newly formed independent national body, the NHS Commissioning Board – now NHS England – in the reforms in June 2012).\(^{34}\)

The social insurance program administered by the government in the USA, have introduced an initiative designed to improve patient safety whereby hospitals are not fully reimbursed if the patient is affected by certain patient safety events.\(^{38}\) In England, in 2010 the government announced a policy to levy financial penalties on hospitals for care that results in an emergency readmission. As the subject of these policies, this highlights the importance of adverse events to hospitals. This rise in national action may also have been driven by the increased public and media attention, which is discussed below. This is coupled with the increased focus on the health services in general due to, for example, the problems associated with aging populations, rising costs, lack of accountability and inequalities.\(^{32}\)

#### 1.3.1.2. Intra-governmental priority

Another possible driver for the raised priority is interest taken by international organisations. For example, in 2004 the WHO, following pressure that the Organisation and its member states paid closer attention to patient safety, set up the World Alliance for Patient Safety.

### 1.3.2. Hospital focus

#### 1.3.2.1. Importance to hospitals

The additional length of stay and procedures outlined above have an operational cost to hospitals due to the burden on their resources, both in terms of capacity and workforce. For example, a 2007 study conducted in Japanese hospitals estimated there were 6,240 person-hours dedicated to patient safety practices per 100-beds each year, including about 1,141
person-hours for infection control alone.\textsuperscript{39} The study included incident reporting, external audit, management of medication and maintenance of medical equipment. It did, however, not include the staff hours associated with the additional lengths of stay and procedures. Moreover, the increasingly common policy of patient choice, whereby the person is able to select who provides their healthcare, means any hospital with a reported poor record on adverse events is likely to miss out on possible treatment opportunities and the funding attached to them.

\textbf{1.3.2.2. Importance of acute care}

Acute care is an important part of any healthcare service. Around a third of personal health care expenditure in the USA is committed towards hospital care, with this proportion having increased recently.\textsuperscript{40} In England, in 2012-13, half of local health spend was on “general and acute”, “accident and emergency” and “maternity” (£47 billion out of £91 billion), which is more than double that spent on primary care (£21 billion on GP Services, community prescribing costs, dental services).\textsuperscript{41} Moreover, the estimates given earlier on the prevalence of patient safety events are predominantly based on the hospital setting and, as such, it’s a known area of importance.

In particular, the acuity means the likelihood of a perceptible change following an adverse event is important as it aids outcome measurement whilst also reiterating the high impact (and therefore the need to address it).

\textbf{1.3.3. Patient and media focus}

\textbf{1.3.3.1. High profile cases}

A number of high-profile patient safety events have raised the public’s awareness of medical errors. In England, the death of a child at Bristol Royal Infirmary following surgery - which was undertaken against the advice of anaesthetists, some surgeons and the Department of Health (DH) - was widely reported in the media and led to an external inquiry. A subsequent investigation into the cases of 53 children treated at the hospital, led to three doctors being found guilty of serious professional misconduct. Other countries have had similar cases capture the media and public’s attention. For example, the case of Willie King in Florida, who had the ‘wrong leg’ removed, and that of Libby Zion, who died following the reaction of a drug administered by an overworked, unsupervised junior doctor, highlighted patient safety issues in the US.\textsuperscript{36}
1.3.3.2. Emerging areas of focus

The interest of patients and the media has been further intensified by the emergence of healthcare associated infections as an issue in health service delivery. Studies into the prevalence of infections, combined with high-profile outbreaks, have focused this interest. Both Methicillin-resistant Staphylococcus aureus (MRSA) and Clostridium difficile (C-diff) have emerged as serious issues in healthcare provision and are increasingly seen as amenable to better healthcare.

1.3.3.3. Increased awareness

Further to the media coverage of high-profile incidents, the prevalence means that many people have personal experience of adverse events. A survey conducted in the US found that around 42% of people reported a medical error having affected themselves, a friend or relative. As a result, there is heightened concern about such errors, with respondents to the survey reporting the health care system as ‘moderately safe’. Another survey, conducted by the American Society of Health-System Pharmacists, found that over half of Americans are ‘very concerned’ about either being given the wrong medicine (61%) or complications from a medical procedure (56%).

1.4. Need for indicators

Given the scale of the issues as outlined above, there is a clear need for measurement efforts to help understand prevalence, causes and preventability. In particular the demand comes from a range of sources, as outlined below.

1.4.1. Indicators

1.4.1.1. History of measurement

While the focus on measurement has intensified, it is not new; such studies can be dated to the first anaesthetic death in 1848, or perhaps even earlier with Florence Nightingale reporting in the mid-1800s that patients in London had a greater risk of dying if hospitalised than if cared for at home. The increased focus on quality and safety was not instantaneously matched by additional measurement efforts. As expressed by Mainz (2003), there is a need for measurement efforts to catch up with the interest in quality and safety:

“Much of the interest in quality of care has developed in response to recent dramatic transformations of health care systems, accompanied by new organisational structures and reimbursement strategies that may affect the quality of care. However, only of late has systematic evidence about quality of care begun to be collected in most health care systems.”
1.4.1.2. Definition of indicator

Indicators of the quality and safety of healthcare have been defined in several ways.\textsuperscript{31} For the purpose of this project, I use the definition based on that proffered by Joint Commission on Accreditation of Healthcare Organizations (JCAHO),\textsuperscript{48} as set out below. The requirements of a good indicator are discussed later (paragraph 3.2.3, p66).

\begin{quote}
\textbf{‘Indicator’} refers to a quantified measure that can be used to monitor and evaluate the quality or safety of important governance, management, clinical, and support functions that affect patient outcomes.
\end{quote}

1.4.1.3. Types of indicators

These indicators can take many different forms in terms of the format of their numerical output and the subject they evaluate. They can measure sentinel events or rate-based incidents and can be related to structure, process or outcome of healthcare (Table 2). I have also added ‘proxy outcome’ which is often ignored in the literature as a possible indicator. Each of the types of indicators has advantages and disadvantages. For example, while outcome measures can evaluate actual health improvements they are susceptible to influence from factors other than the standard of healthcare, such as environmental aspects.

\begin{table}[h]
\centering
\begin{tabular}{|l|l|l|}
\hline
Type of indicator & Detail & Example \\
\hline
Sentinel events & Count of never events & Number of patient who die during surgery \\
Rate-based indicator & Proportion of admissions with preferred treatment or outcome & The proportion of patients in hospital during a study period which acquire a healthcare associated infection \\
Structural & Pertaining to organisational resources & Evidence that hospital uses specialist wards \\
Process & Access to specific treatments & Proportion of trauma patients receiving urgent scans \\
Outcome & Relating to patient’s health following intervention & Mortality or emergency readmission rates \\
Proxy outcome & Treatment suggesting a certain outcome & Patients receiving blood transfusions following treatment (suggesting post-operative haemorrhage) \\
Intermediate outcome & Biological indicator associated with higher risk of poor outcome & Proportion of patients with high blood pressure \\
\hline
\end{tabular}
\caption{Examples of different indicators}
\end{table}

Source: adapted from Table 2 and 3 in Mainz 2003\textsuperscript{33}
1.4.1. Indicator bundles

While inherently interesting, outcome measures have been criticised for their reliance on appropriate risk-adjustment models to account for case-mix, with suggestions that – for the purpose of monitoring hospital performance – non-mortality outcomes or process measures should be developed. As far back at 1992, there was a recognition that most efforts had concentrated primarily on mortality rates.

Of course these types of measures can be, and have been, used in conjunction with one another with, for instance, one US hospital using a: “framework to monitoring patient safety, combining a valid rate-based measures to evaluate outcomes and processes of case, and non-rate-based measures to evaluate structure and context of care.” Indeed, in a review of methods for measuring errors and adverse event endorsed a comprehensive monitoring system using a combination of methods. The advantages of applying a range of measures is discussed later (paragraph 2.5.2, p52).

1.4.2. Demand for indicators

1.4.2.1. Under-reporting

McIntyre and Popper noted in 1983 that “errors need to be recorded and analysed if we are to discover why they occurred and how they could be prevented”; however, emerging evidence suggests that there is considerable underreporting from the current systems. For instance, a study found that in England, only 7% of patient safety incidents identified using a case note review were included in the routine reporting system, with an even lower positive predictive value (5%) for incidents resulting in harm. Whilst not at such a high level, a review of underreporting by Douglas and Pronovost presented findings from studies from across the USA and UK that showed a majority of errors go unreported, with evidence that more serious errors are often not reported. The authors conclude that this underreporting has serious consequences: plaguing the knowledge of the issues; skewing interest toward those events that are reported; and leading to an inability to generalise findings.

1.4.2.2. Increasing demand for measures

As a result of this increased focus, all these stakeholders – from the public to governments and international organisations – have demanded the tools to help analyse the problem and to facilitate improvements. There has been a simultaneous increase in importance to healthcare providers, commissioners, regulators, and patients of assessing the quality of
The increased focus on quality improvements – related to, for instance, the aging populations, higher prevalence of chronic diseases, the introduction of new medical technologies, and increased public expectation\cite{57,58} – has been coupled with an increased demand for measures of quality and safety.

### 1.4.2.3. Measurement efforts remain short

Yet the research to date has failed to satisfy the need; “the literature [on measuring adverse events] consists mostly of descriptive studies. Articles on many of the topics of interest are rare, and the studies are of variable quality” (p 10).\cite{59} Whilst studies have given an indication of the scale of the problem, without better measurement some aspects will inevitably be ignored and go unaddressed with quality and patient safety failings remaining the elephant in the room.

### 1.4.2.4. Government and intra-governmental demand for measures

Such measures are also needed as the foundation of some governmental policies. The requirement is particularly conspicuous in the healthcare systems which have implemented a patient choice agenda, such as the NHS in England. Further to fitting in with current strategies, Governments also require this information to set future priorities:

“Greater priority should be given to clinical areas where there is evidence that the quality of care is either variable or substandard, so that areas with a substantial potential for quality improvement are chosen.” (p16)\cite{60}

And the move towards measuring quality of care has not only been a function of Governments’ requirements, rather the lead has been taken by international organisations, such as the WHO and Organisation for Economic Cooperation and Development (OECD).\cite{32}

### 1.4.2.5. Increase in guidelines

Many indicators – especially process measures – are based on clinical guidelines. Where there is sufficient clinical evidence, these guidelines are based on a review of these existing findings or elsewhere the standards are based on a formal process to gain appropriate clinical consensus.\cite{33}

There has been a proliferation in clinical guidelines. For instance, in England, the National Institute for Health and Clinical Excellence (NICE) which was only formed in 1999 had, by 2013, 165 clinical guidelines (coupled with guidance for 381 interventional procedures, 275 technology appraisals and 42 public health areas).\cite{61}
1.4.2.6. Demand for new approaches

As well as new measures (paragraph 1.4.2.2, p32) a more general demand has been articulated for new approaches to addressing adverse events. The implication from a number of authors is that measures have to be further developed: "Efforts to improve patient safety have fallen short of IoM’s ambitious goal" (p850) and "the overall approach to patient safety...and definitional issues... remains debated" (p1868). Indeed the message that the current approach is inadequate has also been echoed in the UK. In particular, there has been a demand for more systematic and on-going methods:

“Adverse events are systematically and frequently reported in European countries and analyses are made in order to learn from incidents and improve safety. This is resource consuming work, which creates a major need to measure dimensions of safety on an ongoing systematic basis, implement learning organisations, demonstrate ongoing safety improvement, determine when lapses in patient safety occur, and document positive effect of the efforts made.” (p4)

1.4.3. Strengths and weaknesses of methods for identifying and analysing adverse events

A number of different techniques have been used for measuring adverse events, including: review of medical records; interviews with healthcare providers; direct observation; active clinical surveillance; morbidity and mortality conferences; incident reporting system; external audit and confidential enquiry; studies on claims and complaints; autopsy reports; IT and electronic medical records; and administrative data analysis.

The advantages and disadvantages of the different methods are summarised in Table 3. In 2003, two reviews were published on the relatively advantages and disadvantages of methods for assessing patient safety. With other, more recent, comparisons of strengths and weaknesses having been published, I do not restate these in depth here.
### Table 3: The strengths and weaknesses of methods for measuring quality and safety

<table>
<thead>
<tr>
<th>Category</th>
<th>Method</th>
<th>Strengths</th>
<th>Weaknesses</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Ad hoc studies on epidemiology designs and systematic data collections</strong></td>
<td>Review of medical records (case note review)</td>
<td>Uses readily available data&lt;br&gt;Commonly used&lt;br&gt;High reliability of extracted data</td>
<td>Judgments about adverse events not reliable&lt;br&gt;Medical records are incomplete&lt;br&gt;Hindsight bias&lt;br&gt;Format not readily usable for research</td>
</tr>
<tr>
<td></td>
<td>Interviews with healthcare providers</td>
<td>Potentially provides additional data on root causes</td>
<td>Hindsight bias*&lt;br&gt;Reporting bias*</td>
</tr>
<tr>
<td></td>
<td>Direct observation</td>
<td>Potentially accurate and precise&lt;br&gt;Provides data otherwise unavailable&lt;br&gt;Detects more active errors than other methods</td>
<td>Time consuming and expensive&lt;br&gt;Difficult to train reliable observers&lt;br&gt;Potential concerns about confidentiality&lt;br&gt;Possible to be overwhelmed with information</td>
</tr>
<tr>
<td></td>
<td>Active clinical surveillance</td>
<td>Potentially accurate and precise for adverse events</td>
<td>Time consuming and expensive</td>
</tr>
<tr>
<td></td>
<td>Morbidity and mortality conferences</td>
<td>Can suggest contributory factors&lt;br&gt;Familiar to healthcare providers</td>
<td>Hindsight bias&lt;br&gt;Reporting bias&lt;br&gt;Focused on diagnostic errors&lt;br&gt;Infrequently used</td>
</tr>
<tr>
<td></td>
<td>Incident reporting system</td>
<td>Provide multiple perspectives over time&lt;br&gt;Can be a part of routine operations</td>
<td>Reporting bias&lt;br&gt;Hindsight bias&lt;br&gt;Data often unavailable for research</td>
</tr>
<tr>
<td></td>
<td>External audit and confidential enquiry</td>
<td>Provides mechanism for identifying root causes&lt;br&gt;Integrates multiple data sources</td>
<td>Validity of the data untested&lt;br&gt;Cost for using data uncertain</td>
</tr>
<tr>
<td></td>
<td>Studies on claims and complaints</td>
<td>Provides multiple perspectives (patients, providers, lawyers)</td>
<td>Hindsight bias&lt;br&gt;Reporting bias&lt;br&gt;Non-standard source of data</td>
</tr>
<tr>
<td></td>
<td>Autopsy reports</td>
<td>Provides data otherwise unavailable (e.g. on missed-diagnoses)&lt;br&gt;High reliability of data</td>
<td>Biased data as not systematically undertaken</td>
</tr>
<tr>
<td></td>
<td>IT and electronic medical records</td>
<td>Inexpensive after initial investment&lt;br&gt;Monitors in real time&lt;br&gt;Integrates multiple data sources</td>
<td>Susceptible to programming and/or data entry errors&lt;br&gt;Expensive to implement</td>
</tr>
<tr>
<td></td>
<td>Administrative data analysis</td>
<td>Use readily available data&lt;br&gt;Commonly used</td>
<td>Judgements about adverse events not reliable&lt;br&gt;Medical records are incomplete&lt;br&gt;Hindsight bias</td>
</tr>
</tbody>
</table>

Note: * assessed by author

1.4.4. No consensus on a preferred technique

There is no consensus over which is the best method for measuring adverse events and neither is this surprising, since the utility value of each method depends on the purpose, whether that be identifying rates of incidents, ascertaining causes or acting as a warning system.20

Despite the range of methods available for identifying and analysing adverse events, quality and safety shortcomings remain a significant issue and there is little consensus over how best to address it. A study commissioned by the WHO, completed in 2003, identified 262 articles relating to methods for identifying and analysing adverse events.59 The report evaluated the effectiveness of each of the methods against criteria defined by the WHO which highlighted a lack of a preferred method. This conclusion was echoed in a further study published that year.52 That is not to say that the WHO think that having no preferred method means the pursuit of measuring adverse events is futile; conversely it has called for improved efforts, so “making them visible” (p4).70

1.5. Implications for scope

This chapter has set out the case for addressing the shortcomings in quality and safety and provided the rationale for investigating new approaches, especially those with potential to be active, on-going surveillance tools. The chapter also explained that the study incorporates measures of what would be typically classified as quality measures alongside patient safety indicators due to this difficulty to disaggregate between the two types and since they both have intrinsic interest. The estimates on prevalence point to a lack of consensus on the scale. This suggests value in both further national efforts and exploratory work on their preventability, scale and causes.

The next chapter sets out the case for using indicators based on administrative data which were scored promisingly in the above-mentioned WHO-commissioned review of methods59 to meet this demand for an on-going measurement tool. In particular, the chapter focuses on the potential benefits of applying such indicators to specific specialties.
Chapter 2.
Specialty-specific indicators

Current measurement: Case for specialty-specific use

Stroke
Review of current indicators
Obstetrics

Stroke
Application: cross-sectional comparison
Obstetrics

Stroke
Application: temporal analyses
Obstetrics

Discussion
PART I: Introduction

Overview

"You can’t manage what you can’t measure"
quotation on p1, Weingart (2003)

Context

The previous chapter presented the argument for measuring quality and safety in healthcare and the importance within that of acute care. In this chapter the existing literature is reviewed to identify potential tools for addressing this unmet need.

Methods:

A narrative literature review

Findings

There is a strong rationale for using administrative data, including that it has the potential to meet the demand for active surveillance tools for quality and safety. Furthermore, specialty-level development and application of quality and safety indicators – rather than the organisation-wide measures which dominate the current debate – can overcome some of the known limitations of such indicators.

What this chapter adds

Introduces the concept of specialty-specific indicators
First explicit rationale for using specialty-specific indicators based on administrative data

Related papers

Chapter 2: Specialty-specific indicators

2.1. Use of administrative data

2.1.1. Introduction to administrative data

2.1.1.1. Definition of administrative data

Administrative data potentially encompass a large array of collection methods and output and the term remains ill-defined. These differences in administrative hospital datasets include variation in the fields and specific codes for diagnoses, procedures, and timings.

‘Administrative data’ are, for the purpose of this project, defined using the following inclusion criteria:

- routinely collected data covering more than one hospital;
- used for administrative purposes (e.g. for reimbursement); and
- can be replicated using the internationally-recognised ICD9 or ICD10 diagnoses coding framework

2.1.1.2. Common datasets

The most widely used administrative data in research include those of the Agency for Healthcare Research and Quality’s (AHRQ) Healthcare Costs and Utilization Project (HCUP), in particular:

- the State Inpatient Databases (HCUP SID). Currently, 44 States participate in the HCUP SID, encompassing 95% of all USA’s community hospital discharges, with data available from 1990;
- the Nationwide Inpatient Sample (HCUP NIS). The largest all-payer inpatient care database in the USA, containing approximately 8 million hospital stays from around 1,000 hospitals, equating to a 20% stratified sample of USA’s community hospitals, with data available from 1988.²

Within the recent history of research on quality and safety, there has been a rise in the use of administrative data. To demonstrate the use of some key datasets, I conducted a general search of the literature.¹ The results give an indication of the rise in use of such datasets and, while the

¹ I performed a search of title, abstract and key words fields within Medline (Ovid MEDLINE(R) In-Process & Other Non-Indexed Citations and Ovid MEDLINE(R) 1946 to Present), limited to the names (and common abbreviations or alternate spellings) of key datasets in the USA and, for England, Hospital Episode Statistics.
PART I: Introduction

limitations of the search strategy render the results imprecise, the conclusion that there has been an increase in the publications based on administrative data is unquestionable (Figure 1).

Figure 1. Trend in publications based on administrative datasets

Notes: USA total is a sum of counts for the three non-HES datasets

2.1.2. Use of data

2.1.2.1. Background in non-healthcare settings

The use of administrative data to assess quality has been pioneered by the aviation, petrochemical and nuclear energy industries with their so-called High Reliability Organisations (HROs). Whilst the use of administrative data to measure healthcare outcomes is not new; the heightened attention being paid to quality and safety in healthcare is increasing the demand for data. Such data were used first, in the 1970s, to show variations in care, and its use developed to outcomes measurements in the 1980s before moving on to quality and patient safety since the 1990s.

2.1.2.2. Growth in indicators

Coinciding with a growth in quality indicators, there has been a rise in the use of administrative data. To exemplify this I searched Medline (Ovid MEDLINE(R) In-Process & Other Non-Indexed Citations and Ovid MEDLINE(R) 1946 to Present) to give an indication of growth in the use of quality indicators and administrative data. I limited the search to the (MESH) indexing term “quality indicator, healthcare” and search term “administrative data” on the title, abstract and key word fields.
Chapter 2: Specialty-specific indicators

Figure 2. Trends in publication of quality indicators and use of administrative data

Administrative data are collected in many countries and there is scope to expand the use of these data to flag issues in the quality and safety of hospital care, both in terms of wider and better coverage of the indicators applied and a broader range of health services subjected to evaluation.

2.1.3. The development of indicators

2.1.3.1. Historical overview of indicator development
Since the development of the original quality indicators in the early 1990s, many medical providers, healthcare organisations and researchers have contributed to their advancement and understanding of their efficacy. A paper in 2001 suggested that the number of quality indicators had risen to over 200, and by 2009 one initiative, the International Quality Improvement Programme, had identified nearly 700. However, these numbers should be interpreted with caution as a small variation on an existing algorithm for a measure could be classified as a new indicator thus inflating overall counts.

2.1.3.2. Key indicator sets
A key set of indicators was proposed by the Complications Screening Program (CSP) for screening for variations in quality in 1992. Since then, the Agency for Healthcare Research and Quality (AHRQ) has been at the forefront of indicator development. Examples of key indicator sets are given in Table 4.
PART I: Introduction

Table 4. Examples of key quality and patient safety indicator sets

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Reference</th>
<th>Example of indicators/groups</th>
<th>Key associated studies</th>
</tr>
</thead>
<tbody>
<tr>
<td>Complication Screening Program</td>
<td>CSP indicators</td>
<td>Iezzoni et al (1992)</td>
<td>27 indicators, e.g.: Septicemia, wound infection, cellulitis or decubitus ulcer.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Lawthers et al (2000)</td>
<td></td>
</tr>
<tr>
<td>HCUP Quality Indicators (1994)</td>
<td>HCUP QIs</td>
<td>Ball et al (1998)</td>
<td>33 indicators, e.g.: Hysterectomy, low birth weight, mortality after knee replacement</td>
</tr>
<tr>
<td>Refined HCUP Quality Indicators</td>
<td>HCUP II QIs</td>
<td>Davies et al (2001)</td>
<td>25 provider-level, e.g.: stroke mortality, VBAC rates, CE volume</td>
</tr>
<tr>
<td>HCUP Patient Safety Indicators</td>
<td>HCUP PSIs</td>
<td>Miller et al (2001)</td>
<td>12 groups, e.g.: transfusion reaction, postoperative infection, obstetric misadventure</td>
</tr>
<tr>
<td>AHRQ Patient Safety Indicators</td>
<td>AHRQ PSIs</td>
<td>McDonald et al (2002)</td>
<td>20 hospital-level indicators, e.g.: decubitus ulcer and birth trauma</td>
</tr>
<tr>
<td>Indicators for Patient Safety, OECD</td>
<td>OECD HCQI</td>
<td>Millar et al (2004)</td>
<td>21 indicators, e.g.: Wrong blood type; ventilator pneumonia; birth trauma</td>
</tr>
<tr>
<td>Safety Improvement for Patients in Europe</td>
<td>SimPatIE</td>
<td>Kristensen et al (2007)</td>
<td>42 indicators with 24 recommended, e.g.: wrong site surgery; complications of anaesthesia; decubitus ulcer</td>
</tr>
</tbody>
</table>

Notes: VBAC - Vaginal Birth After Caesarean; CE – Carotid Endarterectomy

2.1.3.3. Development of indicators

In the editorial introduction to the first major set of quality indicators – CSP, Iezzoni et al (1992) – the commentator states that they “believe the algorithms will prove difficult to improve upon.”\(^{(50)}\) However, there have been a number of improvements in the indicators to address the limitations of the previous iterations. For example, in 2001 AHRQ published a new set of indicators to address some of the shortcomings of some of the original Healthcare Cost and Utilization Project (HCUP) Quality Indicators, such as their lack of: severity or risk adjustment; measures covering non-surgical medical care; and reliability.\(^{(87)}\) Despite these developments, even by 2008, this “valuable source of information on patient activity remain[ed] under-utilised.”\(^{(88)}\)

2.2. Advantages of using administrative data

In a review of available methods for measuring adverse events, Michel highlights that: (1) rapid assessment methods are needed; (2) the relevance of the method depends on the patient safety
measurement goal; and (3) the relevance of a method depends on the availability of the data. The first and third of these points are covered in the advantages set out below, while the second is integral to the discussion about the purpose of the reporting, raised in Chapter 11. There are many advantages offered by using administrative data to evaluate quality and safety, include those listed below. Advantages are, of course, contextual but those set out here can be viewed in the light of surveillance (a purpose whose rationale is set out above).

2.2.1. Longitudinal
Administrative data are typically continuous. This property permits two additional routes of investigation over snapshot cross-sectional studies: tracking the performance of providers over time (multiple cross-sectional studies) or tracking patients over time (cohort studies) by linking of patients’ records from different admissions. This longitudinal nature is particularly useful for identifying, by looking at specific readmissions, an adverse event for which the symptoms do not present until after discharge.

While this project focuses on acute care, it is worth noting that this longitudinal property of the data allows the researcher to make inferences about what happens outside of hospitals, which has been an area of increasing academic interest. As Weingart and colleagues pointed out, little is known outside hospitals, but administrative data have the potential to record, through emergency admissions and readmissions, an adverse event which may occur outside the non-acute setting. Already, patient safety projects are looking to extend the focus of their research to beyond purely inpatient admissions.

2.2.2. Ease and cost-effectiveness of use

2.2.2.1. Accessibility of data
Administrative data are generally collected to support monitoring or reimbursement activities and, therefore, need to be readily accessible. This advantage has been pointed out in the literature and, more directly, one paper concluded that such information represented “the most readily available set of data systems for tracking patient safety” (p9).

2.2.2.2. Computer readable
The monitoring and reimbursement purposes of administrative datasets involve analysis of the information they contain. This property of being in computer-readable and analysis-ready format makes the use of the data for studying adverse events easier.
2.2.2.3. Cost-effective

Similar to the advantages of being accessible and computer readable, an additional benefit stems from the administrative data having already been collected and so can usually be accessed at little or no cost.\textsuperscript{24,33,68,71,73,92} "Administrative datasets provide a substantially cheaper... way of monitoring rates of adverse events"(p24).\textsuperscript{93} International examples have shown that the costs of running nationwide programmes for disseminating information from these indicators are also relatively low. For example, the Danish National Indicator Project (NIP), which monitors clinical indicators (both quality and patient safety indicators) across six diseases, costs approximately US$750 000 nationwide per year. The project team estimate that the marginal cost for adding additional disease areas to the project would be in the region of US$75 000.\textsuperscript{56} In England, the HES database costs an estimated £1 per record to collect compared with around £10 - £60 per record for clinical registers.\textsuperscript{89}

2.2.3. Timely

Since the monitoring and reimbursement needs to be done in a timely way, the administrative data are provided regularly and with limited time lag.\textsuperscript{89} For example, in the NHS in England, the underlying data that are the basis for the Hospital Episode Statistics (HES) administrative dataset are usually updated every month. As such "administrative datasets provide a... more timely way of monitoring rates of adverse event"(p24).\textsuperscript{93}

2.2.4. Coverage

2.2.4.1. Large populations

Another attraction of administrative data are that they covers large populations\textsuperscript{71,89,91} with, in many cases, the information covering every inpatient procedure offered in hospitals. Again taking England as an example, HES include not only care given by NHS hospitals but also care provided to NHS patients by the independent sector and care given to private patients in NHS hospitals. In 2007-08, the HES dataset included more than 15 million episodes of inpatient care and over 54 million outpatient attendances.\textsuperscript{94}

2.2.4.2. Increasingly large proportions of healthcare

Historically, administrative datasets focused predominantly on subsets of inpatient care. However, in response to the demands for increased data to support initiatives such as pay-for-performance within hospitals, there is increasing coverage of services included in administrative data. For example, in the NHS, national data on outpatient episodes and A&E attendances are also recorded, accounting for approximately 12 million and 40 million respectively each year.
2.2.5. Standardised definitions and rules

Administrative data are based on standardised definitions and rules which have been improved through a number of iterations. The first international classification edition was adopted in 1893. The role of updating the framework was passed to the WHO at its creation in 1948, and regulations adopted in 1967 stipulate that Member States use the updated classifications. The 10th revision of International Classification of Diseases (ICD-10) was introduced in 1990 and is used in over 100 countries. However, there are some notable exceptions, with the USA predominantly using an updated ICD-9 framework. These have been developed over a number of years with many iterations and now offer a relatively comprehensive list of codes and associated guidance.

2.2.6. Basis for Pay for Performance schemes

2.2.6.1. Resulting in improved coding

Administrative data are increasingly used for more rigorous performance monitoring and as the basis for reimbursement schemes. Although this does increase the incentive for gaming, it is also a motivation for increased depth of coding, for example coding more secondary diagnoses. Further, the use of administrative data for reimbursement schemes is often associated with an augmented data audit regime, as was the case in the NHS where the Audit Commission have assumed more work in researching coding accuracy. The expectation is, therefore, that the accuracy and depth of coding is improving.

A systematic review, published in 2012, on discharge coding accuracy of routinely collected data in Great Britain found that primary diagnosis accuracy has improved from 73.8% to 96.0% (p=0.02) since the introduction of Payment by Results in 2002. The authors concluded that such data are sufficiently robust for use in research and managerial decision making.

2.2.6.2. Some adverse events can be costed

This reimbursement aspect has other advantages. Given that healthcare systems are resource constrained, a full understanding of cost implications is important for policy makers to choose priorities. The use of administrative data to reimburse hospitals means that, when a possible adverse event is identified, the cost associated with that admission can be readily estimated. As a result, the economic impact of adverse events can be approximated, and hospitals and governments can make more informed decisions on resources or service redesign.

In 2014, legislation to abolish the Audit Commission was passed as law, with a planned closure date of 1 April 2015. However, the organisations involved in setting the Payment by Results framework (Monitor, NHS England and Department of Health) will all have an invested interest in ensuring good quality data.
2.2.7. Independent reporting

The individuals who code the medical records are independent from the team involved in the healthcare procedures. This arrangement has some limitations, such as the lack of knowledge of the care provided increasing the chance of miscoding, which are discussed below. However, whilst this does not mean that the reporting is entirely independent since the clinicians are still involved in recording the procedures on the medical record, indicators based on such data at least offer some level of independence, thus meeting one of the WHO's criterion for good indicators.97

2.3. Limitation of PSIs

“The sensual eye is just like the palm of the hand. 
The palm has not the means of veering the whole of the beast.”

(The Elephant in the dark, Rumi, 1207-1273)

Whilst indicators of quality and safety based on administrative data offer a number of intrinsic advantages as already discussed, some limitations have been raised. These concerns are often linked to the information not being specifically designed to analyse patient safety. The most prominently cited limitations, which constrain the possible applications rather than making administrative data redundant with respect to analysing patient safety, are outlined below.

2.3.1. Attribution of sequence and causation

2.3.1.1. Difficulty to attribute

While administrative data are longitudinal and include dates for procedures, the datasets offer limited information on the timing of events.81,93,98 For instance, it is difficult to accurately deduce the sequence of events, including when a diagnosis was made,68 and therefore doubt can remain over whether harm was related to quality of care, an adverse event or a pre-existing morbidity.34,40 The HES dataset in England, for example, does not explicitly distinguish whether a safety incident (flagged by a diagnosis or procedural code) was due to medical error rather than an existing complication which was present on admission to the hospital.

2.3.1.2. Importance of attribution

Reviewing the existing literature reveals conflicting findings on whether this issue on attribution is important. One study investigating the issue of present-on-admission based on 14 selected AHRQ PSIs found that for five indicators the rate would be lower if ‘present on admission’ was accounted for by the administrative data.99 A study, in the USA, evaluating the sensitivity and specificity of some complication codes at flagging patient safety incidents found, by retrospectively reviewing the
case notes, that the flagged condition was present on admission in 13% of surgical cases and 58% of medical cases. Another study suggested one-third of discharges were flagged by PSIs because of a diagnosis already present at admission. However, some of the adverse events that are dismissed as present on admission could be the result of a patient being harmed during a previous admission or the failure of the primary or tertiary care sectors, and therefore still of interest. In response to this problem a number of studies have altered the coding structure so that pre-existing conditions are flagged.

2.3.2. Judgements of responsibility and preventability

As well as lacking detailed information on the sequences and causation of an adverse event, neither does administrative data contain any consistent approach for recording whether the harm was preventable or rather a complication of care. Although some of the diagnosis codes do explicitly imply preventability and certain adverse events lend themselves to detection through appropriate algorithms, many indicators rely on judgements about whether a sequence of procedures and diagnoses is likely to identify an event which was preventable. This is particularly relevant to indicators of hospital acquired infections, since it may not be clear that the appearance of an infection-related diagnosis code during the admission is the result of a delayed diagnosis of an existing infection or one that has actually been hospital-acquired.

2.3.3. Problems in coding details

The administrative data are based on the coding of case notes using a consistent framework. Further to imperfect information about preventability there are also limitations since the coding frameworks were not developed to specifically analyse quality and patient safety issues and are not comprehensive. As a result, the resultant indicators do not represent a comprehensively exhaustive list of measures as analysing administrative data cannot pick up all possible adverse events. As such, the method would be inappropriate for trying to estimate prevalence. It should be noted that ICD-10 has a generally more detailed coding framework than the predecessor ICD-9 – e.g. viral hepatitis expanded from ICD-9 code ‘070’ to ICD-10 codes ‘B15’–’B19’, and from one to five 3-digit categories so such criticisms raised on the former framework – such as by the Agency for Healthcare Research and Quality – cannot necessarily be generalised to the more comprehensive ICD-10 framework.

2.3.4. Case-mix adjustment

There are also limitations to the extent in which administrative data permit you to adjust for case mix. For example, there are a heterogeneity in severity within many of the diagnosis codes, which is a particular issue within medical or psychiatric specialties. Similarly, it is also a problem
where patients often have complex comorbidities.\textsuperscript{98,99} In some instances there is no effort made to adjust for case mix, for example the SimPatIE project catalogues ‘death in low mortality DRG [disease resource group]’ as not requiring any risk-adjustment.\textsuperscript{104} Risk adjustment is clearly important and, without it, some of the indicators may be redundant:

"the level of the PSIs which monitor harm [e.g. rates of mortality and infections] is usually sensitive to bias caused by the severity of the disease, comorbidities and lifestyle factors. Validity of these indicators is therefore depend on simultaneous collection of patient-related data appropriate for risk adjustment. Empirical evidence suggests that administrative data are not satisfactory in this respect especially when used at clinical (hospital) level"(p14).\textsuperscript{65}

This assessment by Kristensen and colleagues paints a particularly negative picture and is not backed up with any empirical evidence on the effect of such bias.

2.3.5. Accuracy of coding

2.3.5.1. Errors and gaming
Administrative data are collected through the coding of medical records which are the notes made by clinicians on the care given to patients. Therefore the accuracy of the dataset depends on the standard of documentation and accuracy of coding. Also, the primary uses of the administrative data for monitoring and reimbursement might provide a potential incentives for ‘gaming’\textsuperscript{34} and make it "unreliable because of coding problems"(p19).\textsuperscript{105} The issues on coding was raised as far back at 1994 by Romano and Mark\textsuperscript{106} and has been reiterated often since.\textsuperscript{83,99}

2.3.5.2. Hospital variation
Further to general issues of coding, there is evidence of significant differences in the standard of coding within and across hospitals.\textsuperscript{10,81,107} The same study that identified the factors influencing coding standards found that HRG errors – i.e. the degree to which the correct payment was made – ranged across NHS hospital trusts from 0.3 to 52\%.\textsuperscript{108} Factors that might affect the standard of coding in different hospitals include the depth of coding used (there is considerable variation by hospital in the mean number of codes used per record),\textsuperscript{93} level of clinicians involvement, and standards of validation.\textsuperscript{108}

2.3.6. Some have not been applied to ICD-10
As the older ICD-9 coding framework is replaced with the revised ICD-10 system, likewise the algorithms used to flag quality and safety events need to be updated. A group of researchers and users of health administrative data from Canada, the United States, Switzerland, Australia, China and...
the United Kingdom came together in June 2005, and identified the translation of indicators to ICD-10 as one of the high-priority methodological research areas.¹⁰⁹

2.3.7. Limited evidence on effectiveness
There is limited evidence on whether application of such indicators is effective.⁶⁷ This may be a combination of lack of time lag to evaluate this relatively recent development, and difficulties to account for bias created by changing data recording practices over time. However, some of the positive findings to date are drawn out in the discussion (Chapter 11).

2.4. Review of validity of indicators

2.4.1. Defining validity of indicators
Validity is a measure of the effectiveness of indicators. In the project to refine the original AHRQ indicators (HCUP II QI), the following criteria were used to judge the validity of the indicators (Figure 3).⁷⁴ These aspects of validity are not comprehensive with, for instance, no assessment of content validity which refers to the extent to which a measure represents all facets of quality or safety being assessed. Moreover, these dimensions of validity do not necessarily carry equal weight, with this being a function of the method being applied. For instance, “reliability is important when using an indicator to make comparisons among groups or within groups over time” (p524).³³ A more detailed assessment of validity is given in the following Chapter.
2.4.2. Previous literature on validity

2.4.2.1. Examples of validation results

Findings from specific studies on validity are not necessarily generalisable due to, for instance: most assessments covering only a subset of the dimensions of validity; differences in data accuracy and completeness between different datasets. Most evaluation have focused on patient safety measures, presumably as the events being investigated are, possibly more significant by nature, often more identifiable than quality differences so making validation easier. A review of literature found that validation of existing indicators had been largely restricted to the USA and the validity of many individual indicators remains unresolved. Of these publications on aspects of validity, the results are mixed in terms of techniques used and the resultant findings, for example:

- by comparing two reporting systems and conducting a case note review, one study found that one of the AHRQ PSIs (PSI-7, selected infections) performed poorly;

- another study compared an indicator based on administrative data to case note review and found that the former identified 23.3 decubitus ulcers per 1000 patients, to a true rate of 6.14, a 74% variance; and

- in the UK, out of 2,150 medical records that were flagged as having an adverse event by a PSIs, 72.5% were found to match the (translated) AHRQ PSI specification.
2.4.2.2. Difficulty in generalising

Different indicators are susceptible to different bias or errors and their validities will vary considerably. The possible exception is face validity, which is less dependent on the actual dataset being employed and, in this case, Rosen and colleagues found that several studies had demonstrated that such indicators have reasonable face validity.\(^{113}\) Despite this difficulty in generalising findings, some academics inadvisably made sweeping conclusions. A symptom of the issue is the contradictory nature of some conclusions, with some stating that indicators have demonstrated good reliability, based on the consistent results found across studies \(^{113,114}\) and across years \(^{73,115}\) whereas, for instance, Naessens and colleagues concluded that “in general, many patient safety indicators do not reliably identify adverse hospital events” (p781).\(^{101}\)

2.5. Potential applications of indicators

Whilst there are certainly some limitations to the current array of QIs and PSIs, these failings do not rule such indicators either irrelevant or unusable. Rather the floors just limit the extent to which the indicators can be appropriately used. Rivard, Rosen et al noted that “to address the question of how best to use patient safety indicators, one needs to assess both what the indicators can validly be used for as well as what purpose people may want to derive from them”\(^{73}\).

2.5.1. Types of reporting

2.5.1.1. Framework for describing reporting

In general, performance indicators can be used for four different functions: facilitating accountability; monitoring performance; as a learning tool; and forming policy initiatives.\(^{58}\) In exploring the purpose of indicators in comparing performance of NHS hospital trusts, one paper categorised utilisations of indicators through the type of information they can provide: either ‘hard’ (formal and quantitative) or ‘soft’ (informal and subjective) information.\(^{116}\) This is covered in more detail at the end of the thesis.

2.5.1.2. Recommended reporting type

The WHO has recommended that reporting systems should be blame-free, concluding that successful systems are: non-punitive; confidential; independent; analysed by experts; credible; timely; systems-orientated; and responsive.\(^{97}\) This conclusion was echoed, in part, by the team constructing a national set of indicators in Denmark which called for reporting systems to be: confidential, sanction-free, learning oriented.\(^{117}\)
2.5.1.3. Previously recommended applications

As well as setting conceptual frameworks on the use of such indicators, the literature also contains many examples of specific suggestions for these indicators (Figure 4). For example, one paper suggested that “their simplicity and reliability make them valuable as a higher-level safety performance measure. They offer one means for coordination and integration of patient safety data and activity within and across organizations” (p1633).

Figure 4. Suggested indicator applications


2.5.2. Applying sets of indicators

The case for bundles of indicators was introduced earlier (paragraph 1.4.1.1, p32) with specific advantages set out below. An explanation of the domains of validity is given in paragraph 3.2.1.2, p64.

2.5.2.1. Improving content validity

To draw on the analogy of the blind men and the elephant, the story – in some of its interpretations – tells of a number of blind men, each grasping at an elephant but all touching different limbs. The result being that the man holding one of the elephant’s legs makes an entirely different inference of what the object is to his peer grasping its trunk. Whilst none of the men are able to deduce alone the full nature of the elephant, with enough hands and communication between them a correct identification can be made. This holds true with these indicators which cannot alone give a whole
picture of patient safety, however, a range of indicators used in conjunction with other measurement methods could give a useful depiction.

2.5.2.2. Allowing additional validation techniques
Using a set of indicators also provides the opportunity to undertake additional validation. Specifically, the validation is based on the hypotheses that:

- Two similar measures of the same construct should yield similar results (construct validity)
- Process and outcome measures evaluating the same aspect of care should be correlation (convergent validity);
- An indicator should be associated with “better” measures of performance (criterion validity)

2.5.2.3. The Donabedian model revisited
As explained earlier (paragraph 1.2.1, p24), the commonly cited model by Donabedian regarding patient safety incidents describes the interplay between structure, process and outcomes. By using appropriately focused set of indicators, these various factors can all be investigated, to some extent, to help epidemiological study design.

2.5.3. Primarily organisational-wide indicators to date

2.5.3.1. Legacy of broad indicators
To date the focus of performance monitoring using administrative data has been organisation-wide, cross-specialty indicators, in particular hospital standardised mortality rates (HSMRs). This phenomenon was cited in the rationale for developing the first major set of quality indicators (CSP, lezzeno et al 199250) however recent publications have concluded that this limited focus remains the same.49 A key example of the prominence of these broad measures are the well-publicised HSMRs based on routinely collected hospital episode statistics (HES) for the major acute hospitals in England, which have been published since January 2001 and, indeed, the monitoring of organisational-wide indicators has coincided with some improved results, such as decreases in recorded hospital-standardised mortality rates.125

2.5.3.2. Limitations
However, there are substantial limitations to such broad indicators. Given the widely held belief that it is important to report performance indicators to clinicians (as discussed in more detail at the end of this thesis, paragraph 11.5.1, p260) it is surprising that researchers and regulators have persisted in relying predominantly on organisation-wide indicators of performance which have only limited
relevance to an individual clinician. This situation is epitomised by a recent, comprehensive review of PSIs which purposefully excluded studies “primarily on specific disease, diagnoses or treatments” (p14).

2.5.3.3. **Application of indicators to specific cohorts**

There has been some development of indicators for particular cohorts of patients, such as the long-standing AHRQ Paediatric Quality Indicators which have been available in January, 2006, contains indicators that apply to the special characteristics of the paediatric population. Similarly the first AHRQ HCUP indicator set reported mortality for different groups. Similarly, indicator sets not exclusively based on administrative data – such as the Danish NIP – have commonly included some specialty-specific indicators.

2.6. **Advantages of reporting specialty-specific indicators**

The following section draws on both the limitations of administrative data (described in this Chapter) and those generic limitations in existing measurement (described in the previous Chapter) to show that specialty-specific indicators offer the potential to overcome these key challenges to some extent. Specialty-specific measurement is topical: as suggested in 2012, “after recent UK policy developments, considerable attention has been focused upon how clinical specialities measure and report on the quality of care delivered to patients”.

2.6.1. **Introduction to concept of specialty**

2.6.1.1. **Definition of specialty**

‘Specialty’ refers to a branch of medical science, such as urology or paediatrics. For the purpose of this project, formulating strict criteria of the term specialty is not necessary since the findings on feasibility are generalisable to either broader or narrower groupings. The specialties investigated are defined in more detail in the relevant sections.

In reality, the case put forward and empirical findings on the benefits of specialty indicators hold for those indicators pitched at a lower-then-hospital and higher-than-clinician level.

2.6.1.2. **Differentiating categories of specialties**

Specialties can be classified according to whether they are:
- surgical (i.e. an important part of diagnosis of treatment is achieved through major surgical techniques) or medical (related to internal medicine; i.e. important part of diagnosis is never achieved through surgery);

- the characteristics of patients treated (e.g. paediatricians treat children);

- diagnostic (e.g. radiology) or therapeutic (e.g. geriatrics);

- organ-based (e.g. cardiology) or technique-based (e.g. anaesthetics). 127

2.6.2. Introduction to specialty-specific indicators

2.6.2.1. Demand for specialty-specific initiatives

Although few in number, there are specialty-specific systems for addressing patient safety and, where these have been evaluated, some have shown promise. For instance, Wu and colleagues were able to use a specialty-specific system to analyse patient safety issues and, as a result, implement a number of safety initiatives. 128 “They suggest that, in time, specialty systems such as their own must be integrated with other reporting systems, such as risk management, State or national systems, to minimise duplication of effort and reduce the burden on the staff” (p69). 20 In fact, the author of the 1999 IoM report that was integral to increasing the priority of patient safety has suggested that specialty-based, focused reporting programs should be developed by other specialties. 129

2.6.2.2. Previous specialty-specific initiatives

Previous papers on quality and safety indicators have disaggregated the indicators by specialty or diagnosis. In 2006, Mattke and colleagues recommended the use of 12 mental health, 17 cardiac care and 9 diabetes indicators, 130 whilst another paper recommended the use of indicators categorised by diagnosis: stroke, heart failure, schizophrenia, acute gastrointestinal surgery, heart failure, and lung cancer. 56 In the AHRQ HCUP II indicator set, the following areas of healthcare were given specific, separate consideration: cardiovascular, cerebrovascular, other vascular, geriatric, obstetric, paediatric, chronic conditions, diabetes. 60 Similarly, the SimPatIE indicator set included the AHRQ obstetrics indicators and five single indicators for specific diagnoses. 65

2.6.3. Difference between specialties

The following section highlights some key variations in specialties which suggest that cross-specialty indicators have specific limitations and, as a result, disaggregating performance measurement tools into individual specialties carries advantages.
2.6.3.1. Different cultures

Some specialties have more developed safety cultures. This variation can, in part, be traced historically. The British Ministry of Health set up a committee to examine morbidity and mortality during childbirth in the 1930s. Consequently, a local programme was established, involving antenatal clinics, meetings between midwives and family doctors, a new specialist ward and senior hospital clinician post, which reduced the rate of mortality in that area from 9 to 1.7 per 1000 deliveries within 5 years. This area of medicine has continued to pioneer, in relative terms, patient safety. And the importance of this patient safety culture is highlighted elsewhere.

2.6.3.2. Different prevalence

There is no reason to assume that patient safety is an equally important issue, with similar prevalence, across different aspects on healthcare. For example, the effects and causes of adverse events in elective surgical admissions are, by their nature, more easily identified. Evidence suggests that there are significant variations between specialties in the number of adverse events occurring. For example, in the specialty of mental health, analysis of existing PSIs suggests that people with schizophrenia have higher chances of suffering from post-operative respiratory failure, deep vein thrombosis and sepsis. The same study concluded that there is scope for further work to fill in the gaps in knowledge about adverse events within the specialty of mental health.

As a result, indicators are likely to identify more events in some specialties. Analysis using the AHRQ PSI algorithms suggested that 31% of adverse events identified occurred in obstetric hospitalisation although this will be biased to some extent by the selection of indicators.

2.6.3.3. Different impact

Similarly, there is no reason to assume that PSIs are likely to result in equivalent levels of harm between specialties: anaesthetists’ mistakes, for example, may be more serious than those of doctors in other specialties; and in critical care, due to the high number of patients with complex and severe illnesses, the effects of any adverse event are more pronounced. As a result of these variations, it would be preferable to prioritise some specialties and disaggregate between those specialties with different levels of impact from patient safety events.

2.6.3.4. Applicability of coding framework

The framework used for coding medical records to create the administrative dataset on which PSIs are based is not comprehensive (as discussed earlier in §2.3.3). As a result, the framework will be naturally better suited to accurately identify adverse events in some specialties. For instance, a study that conducted a case note review to check the validity of results from applying the AHRQ PSIs to
Chapter 2: Specialty-specific indicators

NHS data found that the highest levels of agreement were found with the indicators relating to obstetric care. Consequently, previous work on developing indicators has focused on certain specialties. For example, a project covering hospitals in Queensland, Australia chose to focus on mental health, stroke, emergency departments, COPD and medication.

2.6.3.5. Maturity of existing tools for monitoring

Further to the example of Wu et al, some specialties already have established tools for monitoring adverse events. “These systems are designed to provide information on specific clinical issues, which can be shared within the professional group” (p68).

In particular, anaesthesia, neonatal and adult intensive care units have established specific systems.

2.6.4. Nature of specialties

This section discusses the properties of specialties, in particular as groups of clinicians, that make specialty-specific measures, in theory, appropriate.

2.6.4.1. Group learning

The patient safety literature has underlined the importance of group learning as well as organisational and individual learning. McIntyre and Popper highlighted that “experiences need to be pooled so that doctors may also learn from the errors of others” (p1919). Specialties offer such a group which already have existing channels, such as through morbidity and mortality conferences, to be analysing these initiatives.

2.6.4.2. Specialty cultures

One of the keys to improving patient safety is to ensure that the health workforce develops the will to address the issues. This was shown empirically by Kline and colleagues who found that a more positive culture of patient safety within hospital units was related to lower incident severity. A review of patient safety culture literature performed in 2007 also found that clinicians are a key factor in the patient safety culture. Reporting to clinicians can, itself, be seen as an initiative for improving the patient safety culture. Further to the importance of clinicians in improving the patient safety culture, the same reviewers suggested that groups of people, termed as “sub cultures”, such as specialty teams, are also important.

2.6.4.3. Understanding epidemiology and proposing changes in practice

Once a patient safety issue has been identified, a suitable response needs to be developed. Some mistakes in healthcare are as a result of complex processes and structures, such as in medication which involves numerous stages. Given these complexities, specialty teams, and clinicians within them, are the appropriate recipients of the information as they can analyse any issue in the
appropriate context. Certainly, there is recognition that addressing patient safety requires in-depth involvement of clinicians. Leggat and colleagues, for instance, point out the importance of the skill and capacity of healthcare workers in delivering safer patient care.\textsuperscript{136} Involving clinicians has been shown to be effective; for example, involving clinical teams in the monitoring of VTE resulted in a number of practical changes to working practices being made.\textsuperscript{139}

\subsection*{2.6.4.4. Narrowed focus}
There are currently gaps in our knowledge about the epidemiology and effects of adverse events. However, it is clear that to improve patient safety these gaps in our understanding need to be addressed. Mayers-Oakes, in a paper which outlined research in breast cancer, diabetes mellitus, upper gastrointestinal and lower intestinal, concluded that this need could be met by narrowing the focus of indicators.\textsuperscript{140} The implication from this conclusion is that concentrating indicators on specialties rather than, say, pan-hospital issues would be beneficial. Certainly, initiatives with less focus have not always been successful. For instance, Weiner and colleagues suggested that hospital-wide initiatives have not always improved patient safety.\textsuperscript{141} This narrowed focus can also help improve interpretability and increase validity by selecting either homogenous patients or areas where less case-mix adjustment is needed.\textsuperscript{65}

\subsection*{2.6.4.5. Developing new and existing indicators}
The increased demand for indicators (1.4.2) that has precipitated from the raised profile of patient safety should be met, at least in part, by developing and validating more indicators.\textsuperscript{73,130} Clinicians and specialty teams are well placed to help address this need due to their knowledge and experience. Further to meeting the need for additional information, developing additional indicators would also:

- reduce the focus on particular diagnoses and whether they are coded. This would reduce any ability and incentive to ‘game’ by not coding accurately or comprehensively; and

- “create the opportunity to test correlations and identify factors from among indicators. This in turn would enable the indicators to depict patient safety on a number of distinct dimensions”\textsuperscript{(p1647-8)}.\textsuperscript{73}

\subsection*{2.6.5. Disadvantages of specialty-specific}
Many indicators will be specific to that specialty and therefore not applicable to other specialties. One implication is therefore that it is difficult to compare results between specialties and create, for example, control groups for evaluations.\textsuperscript{142} One solution is to create other counterfactuals by comparing between patient groups, hospitals or time-periods.
By being more focused, the indicators may be less likely to measure aspects of the treatment and outcomes that are not anticipated; conversely, broader measures have a higher propensity to detect unexpected issues. However, increasing the breadth of indicators will usually make controlling for bias more difficult.

2.7. Research objectives

2.7.1. Hypotheses
The use of hospital-level cross-specialty indicators is relatively well established. However, there is scope to expand the use of administrative data, in terms of: using more focused indicators; applying and validating the measures in novel ways; and introducing them to new healthcare services. As such, the project was designed to test the following hypotheses:

1. A specialty-specific approach to developing indicators can result in a more useful and valid bundle of indicators to identify quality and safety issues;

2. Hospital-level comparisons could be improved through better validation techniques (coding practice, indicator comparison and organisational factors); and

3. The temporal nature of administrative data can be used to overcome some of the limitations of such indicators and highlight important quality and safety issues.

2.7.2. Thesis outline
To test these hypotheses, a range of qualitative and quantitative methods were applied (Table 5).

<table>
<thead>
<tr>
<th>Hypothesis</th>
<th>Chapter (Key method)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. A specialty-specific approach to developing indicators can result in</td>
<td>Chapters 4 and 8</td>
</tr>
<tr>
<td>a more useful and valid bundle of indicators to identify quality and</td>
<td>(Systematic reviews)</td>
</tr>
<tr>
<td>safety issues</td>
<td></td>
</tr>
<tr>
<td>2. Hospital-level comparisons could be improved through more robust</td>
<td>Chapters 5 and 9</td>
</tr>
<tr>
<td>validation techniques (coding practice, indicator comparison and</td>
<td>(Logistic and multi-level regression modelling)</td>
</tr>
<tr>
<td>organisational factors)</td>
<td></td>
</tr>
<tr>
<td>3. The temporal nature of administrative data can be used to overcome</td>
<td>Chapters 6, 7 and 10</td>
</tr>
<tr>
<td>some of the limitations of such indicators and highlight important quality</td>
<td>(Logistic and multi-level linear modelling)</td>
</tr>
<tr>
<td>and safety issues</td>
<td></td>
</tr>
</tbody>
</table>
PART I: Introduction

Chapter 3.

Project design and methods
Chapter 3: Project design and methods

Overview

Context
The previous two chapters have set out the case for further work in evaluating the feasibility of using specialty-specific indicators based on administrative data. This chapter presents the analytical approach taken. The design of the individual studies on the two specialties is also driven by the specific literature reviews (Chapters 4 and 7) and the implications on the methods are described in these later specialty-related sections.

Methods
Review of existing analytical approaches to developing and utilising indicators based on administrative data.

Findings
Hospital stroke care and obstetrics are appropriate specialties to use as case studies – due to their importance, evidence of scope to improve, and relative differences – for demonstrating feasibility of applying specialty-specific indicators. A range of methods have been used to identify, develop and present indicators.

What this chapter adds
- Set out the analytical opportunity for this new concept of specialty-specific indicators
- Present a new hybrid approach to evaluating indicators
3.1. Introduction to study design

3.1.1. Introduction to feasibility

Now that the case for specialty-specific indicators based on administrative data has been made, the next step is to ensure that they can be applied in a robust fashion. “It is imperative that clinical indicators are meaningful, scientifically sound, generalisable, and interpretable. To achieve this, clinical indicators must be developed, tested, and implemented with scientific rigour”\(^\text{(p15)}\). This project investigates the feasibility of using specialty-specific indicators based on administrative data to meet this need.

‘Feasibility’ is defined, for this project, in broad terms including not only the various dimensions of validity but also the ability to apply these indicators. It is tested here using a reactive approach, namely applying analytical tests to address the gaps in current knowledge (such as assessing the strength of correlations between indicators as a marker of validity) while drawing on previous literature where this is generalisable (such as comparing findings from the indicators to similar research).

3.1.2. Addressing limitation in indicator use

The previous chapter covered the literature on how indicators based on administrative data have been recommended for use, which is covered again in the final chapter. If one accepts the premises that (a) administrative data hold the potential to improve the measurement of quality and safety and (b) the misapplication of administrative data would be a retrograde step, then an implication is that it would be worthwhile to focus the application on mitigating the extant limitations. As such, the studies contained in this monograph are designed to show how to best overcome or evaluate the key limitations from the existing literature, as outlined in Table 6.
### Table 6. Recommended study design to mitigate limitations

<table>
<thead>
<tr>
<th>Para.</th>
<th>Issue</th>
<th>Detail</th>
<th>Recommendation on study design</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.3.1</td>
<td>Causation</td>
<td>Difficulty in attribution of sequence and causation</td>
<td>Range of indicator types: Complement outcome measures with process indicators that inherently are more amenable to hospital performance. Regression analysis between outcomes, activity and structural factors: Investigate associations between hospital performance and other hospital factors.</td>
</tr>
<tr>
<td>2.3.2</td>
<td>Judging preventability</td>
<td>Difficulty to assess responsibility and preventability</td>
<td>Using comparative analysis: Evaluate differences between provider performance and longitudinal trends, complemented with statistical analyses, to estimate achievable performance.</td>
</tr>
<tr>
<td>2.3.3</td>
<td>Coding detail</td>
<td>Problems in coding framework not identifying key events</td>
<td>Systematic review of existing measures: Comprehensive evaluation of what indicators exist and which aspects of care can be evaluated.</td>
</tr>
<tr>
<td>2.3.4</td>
<td>Risk model</td>
<td>Need for case-mix adjustment</td>
<td>Case-mix adjustment: Include detail on the risk model Specialty-level: Limit to less heterogeneous groups of patients</td>
</tr>
<tr>
<td>2.3.5</td>
<td>Coding errors</td>
<td>Variation in coding practice and accuracy</td>
<td>Sensitivity analyses: Re-run analysis to evaluate potential effect of differences in coding practice and accuracy</td>
</tr>
<tr>
<td>2.4.2</td>
<td>USA-bias</td>
<td>Most of the studies originated from the USA using ICD-9</td>
<td>Apply to ICD-10 dataset: Use a non-USA dataset based on the current ICD-10 coding framework</td>
</tr>
</tbody>
</table>

In summary, the recommendations above are that the individual studies’ elements: focus on specialties; apply a bundle of indicators including outcomes and processes; investigate correlations between performance and organisational factors; compare performance between hospitals and over time; outline the case-mix adjustment explicitly; and conduct sensitivity analyses to understand effect of variation in coding practice and accuracy. The specific analytical framework used – which encompasses the addressing of existing shortcomings – is described below.

### 3.2. Analytical framework

#### 3.2.1. Introduction

#### 3.2.1.1. Issues with validation

It is not possible to have a perfectly validated indicator; even if an indicator’s performance was compared to an absolute gold standard then this would be snapshot and not necessarily provide
assurance to all time periods. Some domains of validity “are easier to tap than others; unfortunately, the most challenging domains to tap are often the most useful for health care providers who wish to understand the meaning of the data.”(p4)\textsuperscript{145} In particular, lack of a gold standard means that ‘criterion validity’ is not possible. The domains of validity described here are not ordered in a hierarchical fashion; indeed, it is not possible to rank the relative importance of the criteria. Across the domains there are inherent ambiguities and little consensus on their meaning.\textsuperscript{142}

Concepts of validity are poorly described in literature on patient outcomes, this is not due to a lack of theoretical discussions but a lack of consistency.\textsuperscript{142} I have therefore formulated an analytical framework to evaluate validity and other aspects of feasibility using literature on psychometric testing and patient-based outcome measures and present an interpretation below.

The framework does not represent a comprehensively exhaustive test and, indeed, it would be inefficient to apply every method suggested for taking assurance on validity. In the editorial introduction to the first set of indicators – CSP, lezzoni et al (1992) – the consensual approach involving a panel of physicians to provide face validity and precision, was questioned as not being replicable. In particular, the physicians did not agree on their interpretation of the presence or not of an adverse event within the medical note abstracts when trying to validate measure results on CSP.\textsuperscript{50}

### 3.2.1.2. Aspects of validation

As discussed at the start of this chapter (paragraph 3.1.1), a key aspect of feasibility is to ensure the indicators are valid (the other aspects of feasibility – and how they are considered – are given later in this chapter). Validity has a number of domains, although these are neither comprehensively exhaustive nor mutually exclusive (Table 7). To organise the plethora of concepts, I have grouped the methods for demonstrating validity (individual ‘components’) within four global domains (construct, content, face, and criterion).
### Table 7. Domains of validity

<table>
<thead>
<tr>
<th>Validation domain ('global definition')</th>
<th>Requirement</th>
<th>Validation methods ('component definition')</th>
</tr>
</thead>
</table>
| Construct validity                     | Two measures of the same construct should yield similar results | **Convergent** validity: compare process and outcome measures  
**Discriminant** validity: compare to supposedly unrelated measures  
Comparison to structural factors, e.g. skill mix |
| Content (logical) validity              | Indicator should represent all facets of the subject of evaluation | **Consensual** validity: near consensus of an expert panel (for content) or multi-disciplinary panel (for face validity). These domains are relatively less setting/purpose-specific, so previous evaluations can often be generalised. |
| Face validity                           | Indicator should "look like" it is going to measure what it is supposed to measure | **Concurrent** validity: comparison to previously validated measure  
**Predictive** validity: (similar to concurrent) comparison of predicted score and eventual results from validated measure |
| Criterion (concrete) validity           | Indicator should be associated with “better” measures of performance |                                           |

Source: adapted from concepts described in Romano et al (2007)\(^{145}\) and Brewer (2000)\(^{146}\)

#### 3.2.1.3. Study validation

In addition to the aspects of indicator validity, there are domains of study validity. Key domains are: **ecological** validity (that the study design approximates the real-world); **internal** validity (the extent to which a causal conclusion is correct, involving minimising of systematic error); and **external** validity of study, which relates to how generalisable the results are.\(^{146}\)

#### 3.2.2. Previous indicator validation

Evaluations on how ‘good’ an indicator performs usually involves the notions of ‘validity’ and ‘reliability’; however, these terms are used in many ways in the literature.\(^{147,148}\) Some of this variation in the evaluation of indicators might be appropriate since how ‘good’ an indicator is depends on the context of both the subject and purpose of the measurement. To mitigate this issue of differing contexts, the framework used to evaluate the indicators in this project is based on three previously endorsed frameworks used to evaluate (albeit non-specialty) quality and safety indicators based on administrative data, alongside the domains of validity taken from wider literature:

1. the Danish NIP, where ideal indicators were characterised as being: based on agreed definitions, and described exhaustively and exclusively; highly or optimally specific and sensitive; valid and reliable; discriminatory; relates clearly identifiable events for the user; useful for comparisons; evidence-based.\(^{33}\)
2. the refinement the HCUP QIs, whereby indicators were required to have: face validity, precision, minimum bias, construct validity, ability to foster quality improvement, should or could be used effectively.  

3. OECD HCQI indicator project used: importance (impact on health, policy importance, amenability); scientific soundness (face validity, content validity) and feasibility (data availability, reporting burden).

3.2.3. The analytical framework

3.2.3.1. Methods for evaluation

I synthesised these frameworks into a single hybrid set of criteria that were used to define the project protocol, as listed in the final column (Table 8, with further details on how the specific methods are designed to address this framework given in paragraph 3.6, p77). This is a pragmatic approach built on making efficient use of the data and tools available, since previous efforts – including those of Romano (2007) showed that even a comprehensive assessment can be inconclusive with, for instance, an indicator for obstetric trauma having mixed validity: “Recent evidence on construct validity of this indicator is inconclusive, but the evidence on criterion validity is quite supportive” (p18).
Table 8. Analytical framework

<table>
<thead>
<tr>
<th>Domain</th>
<th>Indicator requirement</th>
<th>Study design</th>
</tr>
</thead>
<tbody>
<tr>
<td>Importance</td>
<td>1. Impact Addresses clear gap, with policy importance</td>
<td><strong>Systematic reviews</strong> to find potentially important applications. Identifying key variations in indicators performance.</td>
</tr>
<tr>
<td></td>
<td>2. Amenable Relates to an outcome that can be influenced by healthcare performance</td>
<td>Comparison between key results and organisational factors.</td>
</tr>
<tr>
<td>Scientific soundness of indicators</td>
<td>3. Face and content validity Sound clinical or empirical rationale for its use, and represent all facets of the subject of evaluations</td>
<td><strong>Literature review</strong> to identify indicators with pre-evaluated face/content validity.</td>
</tr>
<tr>
<td></td>
<td>4. Precision Relatively large variation among providers that is not due to random variation or patient characteristics. <em>Includes responsiveness, discriminative (cross-sectional) and evaluative (longitudinal) validity.</em></td>
<td>Consultation with experts on new measures <em>(consensual validity).</em></td>
</tr>
<tr>
<td></td>
<td>5. Construct validity Supported by evidence of relationship with quality, and related to other similar indicators</td>
<td>Statistical analysis of <em>correlations</em> between the set of indicators, including <em>outcome v. process</em> indicators <em>(convergent validity).</em></td>
</tr>
<tr>
<td></td>
<td>6. Minimum bias Unaffected by systematic differences in case-mix Not incentivise gaming of the indicator</td>
<td>Detailed <em>risk-adjustment model</em> applied to indicators and <em>stratification</em> of key results.</td>
</tr>
<tr>
<td>Applicability</td>
<td>7. Data availability Applicable to administrative data</td>
<td><strong>Application</strong> of indicators to show ICD-10 based data is sufficient.</td>
</tr>
<tr>
<td></td>
<td>8. Reporting burden Minimal reporting burden</td>
<td><strong>Application</strong> of indicators to show that results can be derived without significant burden.</td>
</tr>
<tr>
<td></td>
<td>9. External and ecological validity Generalisability of findings</td>
<td>Explicit statement of <em>assumptions</em> used.</td>
</tr>
</tbody>
</table>

Sources: Adapted concepts presented in Davies et al (2001); 80 Mainz (2003); 33 Millar et al (2004); 85 Romano et al (2007); 145 and Brewer (2000). 146

Iezzoni and colleagues concluded that indicators “needed to be validated if they were to be used; however, what was not obvious was what constitutes validity in this context” (p366). 50 Some aspects of validity and applicability not covered in the evaluations framework – such as the ability of the indicators to be effective and “work well” in a portfolio of quality and safety metrics – are, instead, accounted for in the overall discussion (Chapter 11) since they relate to the reporting system. In particular, “internal validity” (correct causation) is partially proven through tests on convergent validity which may show, for instance, that providers offering more of a certain treatment or with a
different skill mix have better performance but the causation is also appropriate for assessment by clinicians and, therefore, relates to the reporting system.

3.2.3.2. Limitations to evaluation protocol

The key limitation of this validation process stems from no “gold standard” against which the results can be compared (criterion validity). On the precision element, given the inability to know absolutely whether a quality or safety event has occurred, calculating indicator-level false positive and false negative levels - as suggested in some validation frameworks, such as Mainz 2003 and Rosen 2007 - is not possible. However, I have mitigated against this by accounting for hospital-level false positive rates using statistical controls and, where available, drawing on previous research, such as the work by Bottle et al (2008) which found, for instance, that from the 2,150 records audited, 72.5% were found to match the PSI specification.

3.3. Specialty selection

3.3.1. Selection framework

Given the aforementioned case to develop and apply specialty-specific indicators, the first stage of the study was to select specialties. Within the NHS, specialties are not mutually exclusive and there is no single agreed list of recognised specialties. However, given the nature of the research presented here, there was no need for precision in the labelling of specialties. By comparing previous lists of specialties (curriculum list, Royal Colleges of Physicians, Surgery, Paediatricians and Child Health, Obstetrics & Gynaecology, Psychiatrists, Anaesthetists, Radiology, Ophthalmology and the Faculty of Occupational Medicine and College of Emergency Medicine), I created a list of 51 recognised specialties within the NHS, excluding general practice and public health which are not hospital-based specialties.

I took a pragmatic approach to selecting which specialties should be evaluated, starting with the domains of ‘Impact’ and ‘Amenable’ from the evaluation framework (Table 7, p65). These criteria focusing on importance – or ensuring that there is scope for improvement - have been expressed elsewhere. Given the number of specialties, I took a range of additional considerations in selecting the specialties. To increase the generalisability of the findings on applying such indicators, I purposefully chose specialties with differences, particularly in:

1. culture and current initiatives to improve quality and safety, in particular referring to the NPSA which then had research programmes into four specialties: anaesthesia, neonatal, obstetrics and gynaecology, oncology.
2. the range of treatments and outcomes;

3. extent to which the policy of choice of provider is potentially applicable;

4. the extent to which administrative data has previously been applied, including the availability of suitable indicators was considered, although the research plan did allow for the development of new indicators.

I excluded specialties with only rare events since these are less likely to have a sufficient evidence-base to have robust guidelines and, as such, there would be difficulty in knowing which processes might be improved by monitoring. The decision was made in consultation with a public health doctor.

3.3.2. Selected specialties

Based on the framework, I chose to proceed with stroke and obstetrics specialties. This is not the only pair of specialties that comply with the criteria but uniqueness was not a pre-requisite and proceeding with these two are sufficient to meet the goals of the project with any wider assessment would be outside the scope for a doctoral thesis. The differences - and combined coverage of different facets of healthcare - that led to this decision are summarised in Table 9.

<table>
<thead>
<tr>
<th>Criteria</th>
<th>Stroke</th>
<th>Obstetrics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall burden of disease (incidents/year in England)</td>
<td>110,000 stroke(^{151})</td>
<td>700,000 births(^{152})</td>
</tr>
<tr>
<td>Type</td>
<td>Medical</td>
<td>Surgical/Medical</td>
</tr>
<tr>
<td>Status of specialty</td>
<td>New</td>
<td>Longstanding</td>
</tr>
<tr>
<td>Characterisation of measurement</td>
<td>New and predominantly audit based</td>
<td>Longstanding with varied tools</td>
</tr>
<tr>
<td>Indicators in CSP (1992)</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Indicators in HCUP QIs</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Indicators in AHRQ PSIs</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Indicators in Danish NIP(^{144})</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Specific NPSA initiative</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Prevalent</td>
<td>Unknown</td>
<td>Yes</td>
</tr>
<tr>
<td>Evidence of scope for improvement</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Specific patients</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Patient type</td>
<td>Unwell patients</td>
<td>Predominantly well patients</td>
</tr>
<tr>
<td>Choice of provider</td>
<td>Rarely</td>
<td>Potentially</td>
</tr>
</tbody>
</table>

\(^{144}\) The decision was made in consultation with a public health doctor.
3.4. Developing an indicator set

3.4.1. Framework for developing indicators

Jan Mainz, from the Danish NIP, presented a conference paper in 2001\(^6\) – later published in 2003 as a companion piece to an article defining and characterising the NIP measures\(^{144}\) – outlining a method for developing evidence-based clinical indicators. Whilst this framework gives only the shortest consideration of the testing of the indicators by application to the data, it provides a useful guide for the initial stage of the project, and the way in which this project complies with the proposed method is summarised in Table 10.

The OECD indicator set which was published at a similar time used a far less comprehensive method for developing an indicator set, involving an expert panel with each member asked to identify 20 measures they felt had the greatest prospects of being selected. Through a series of conference calls and email discussions, the Patient Safety Panel converged on a final list of 21 indicators.\(^8\)
### Table 10. Development of indicator sets

<table>
<thead>
<tr>
<th>Phase</th>
<th>Steps</th>
<th>Sub-steps</th>
<th>Method applied in this study</th>
<th>Chapters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Planning</td>
<td>1. Choose the clinical area to evaluate</td>
<td>Establish importance (high volume, cost, variation)</td>
<td>Systematic reviews (secondary purpose)</td>
<td>4, 8</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Identify opportunities for clinical intervention</td>
<td>Systematic reviews (secondary purpose)</td>
<td>4, 8</td>
</tr>
<tr>
<td></td>
<td>2. Organise the measurement team</td>
<td>Select group participants</td>
<td>Various consultations/collaborations with clinical specialists, statisticians, and clinical coders</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Organise and divide tasks</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>3. Provide an overview of existing evidence and practice</td>
<td>Present documentation and knowledge from the scientific literature for potential indicators</td>
<td>Systematic reviews (primary purpose)</td>
<td>4, 8</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Establish consensus about existing knowledge and practice</td>
<td>Systematic reviews (primary purpose)</td>
<td>4, 8</td>
</tr>
<tr>
<td>Development phase</td>
<td>4. Select clinical indicators and standards</td>
<td>Select process indicators</td>
<td>Predetermined criteria for selecting indicators, including that both process and outcome measures are used</td>
<td>5, 9</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Select outcome indicators</td>
<td></td>
<td>5, 9</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Identify prognostic factors (risk adjustment)</td>
<td>Systematic reviews (primary purpose)</td>
<td>4, 8</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Establish consensus and rating procedures</td>
<td>Stroke: online consultation Obstetrics: use of previously endorsed indicators</td>
<td>5, 9</td>
</tr>
<tr>
<td></td>
<td>5. Design measure specification</td>
<td>Define indicators and standards</td>
<td>Described in application of indicators for comparing performance. Standards, cohort and risk-model developed through systematic reviews (primary purpose) and testing.</td>
<td>5, 9</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Identify target population</td>
<td></td>
<td>5, 9</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Determine inclusion and exclusion criteria</td>
<td></td>
<td>5, 9</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Devise risk adjustment strategy</td>
<td></td>
<td>5, 9</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Identify data sources</td>
<td>Use of English administrative data</td>
<td>3, 5, 9</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Describe data collection procedures</td>
<td>Process for collecting English data outlined below</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Develop a plan</td>
<td>The analytical plan and techniques are described below</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>6. Perform pilot testing</td>
<td>Application of indicators according to analytical framework (Table 8, p67)</td>
<td></td>
<td>3, 5, 6, 7, 9, 10</td>
</tr>
</tbody>
</table>

Source: Mainz 2003144
3.4.2. Literature review type

3.4.2.1. Meeting requirements of previous indicator development frameworks

The literature review was designed to meet the following development steps set out by Mainz (and as described in Table 10, p71) to: present documentation and knowledge from the scientific literature for potential indicators; establish consensus about existing knowledge and practice; and identify prognostic factors (risk adjustment). To meet these objectives the review would need to be comprehensive and I employed the approach of a systematic review, which are used to collate “all evidence relating to the objectives and to reduce any bias in answering the literature review questions”.

The literature review was also designed to meet the development steps to establish importance and scope for improvement, although these were not a core objective as they do not require such a robust approach.

3.4.2.2. Gaps in specialty literature

As set out above (Table 8, p67, Domain 1) a key component to addressing the feasibility of using specialty-specific indicators is to evaluate the current literature within that specialty and understand what the gaps are. Whilst some of the generic issues with quality and safety indicators are presented in the previous chapter, that discussion also highlighted the variation between specialties which will result in unique advantages and disadvantages for applying administrative data to monitor care.

3.4.2.3. Literature review type

These literature reviews were atypical in that they were not designed to answer a single question – such as the effectiveness of a specific treatment – but rather to understand which, and how, indicators have been applied to date. It was also hoped that the approach would reveal the current shortcomings in the development and utilisation of indicators. Despite the likely heterogeneity of the studies prohibiting formal, quantified meta-analyses, the literature reviews shared the features of a systematic, rather than narrative, review (Table 11). As such, these collations and appraisals of the literature are termed as systematic reviews in this monograph.
### Table 11. Key distinctions between narrative and systematic reviews

<table>
<thead>
<tr>
<th>Core Feature</th>
<th>Narrative Review</th>
<th>Systematic Review</th>
<th>Review protocol</th>
</tr>
</thead>
<tbody>
<tr>
<td>Study question</td>
<td>Often broad in scope.</td>
<td>Often a focused clinical question.</td>
<td>Multiple, specific, non-clinical objectives</td>
</tr>
<tr>
<td>Data sources and search strategy</td>
<td>Which databases were searched and search strategy are not typically provided.</td>
<td>Comprehensive search of many databases as well as the so called grey literature. Explicit search strategy provided.</td>
<td>Comprehensive, reproducible search strategy.</td>
</tr>
<tr>
<td>Article review or appraisal</td>
<td>Variable, depending on who is conducting the review.</td>
<td>Rigorous critical appraisal, typically using a data extraction form.</td>
<td>Extraction form used.</td>
</tr>
<tr>
<td>Study quality</td>
<td>If assessed, may not use formal quality assessment.</td>
<td>Some assessment of quality is almost always included as part of the data extraction process.</td>
<td>Assess quality of both study (e.g. reporting of assumptions) and indicators (e.g. case-mix model).</td>
</tr>
<tr>
<td>Synthesis</td>
<td>Often a qualitative summary.</td>
<td>Quantitative summary (meta-analysis) if the data can be appropriately pooled; qualitative otherwise.</td>
<td>Likely qualitative synthesis due to heterogeneity of study designs and administrative data used.</td>
</tr>
<tr>
<td>Inferences</td>
<td>Sometimes evidence-based.</td>
<td>Usually evidence-based.</td>
<td>Inferences all based on extracted information.</td>
</tr>
</tbody>
</table>

Source: Adapted from West et al (2002)\(^{154}\), Cook et al (1997)\(^{155}\), Green et al (2008)\(^{156}\)

### 3.4.3. Previous systematic reviews

To ensure that the work is not redundant, it is good practice to check whether there are existing or on-going similar reviews. I reviewed existing and on-going reviews by searching the Cochrane Database of Systematic Reviews (CDSR), key websites and making enquiries at key organisations involved in the development, application and validation of indicators. The additional purpose of this process was to both:

- inform the design of the specialty-specific systematic reviews by having a comprehensive understanding of what sources, search terms, inclusion criteria have been used to answer meet similar objectives; and

- use the descriptive results (such as number of indicators collated, dates of studies and countries of origin) from these previous reviews as benchmarks to compare to the results of

---

\(^5\) Other sites searched included: National Institute for Health and Clinical Excellence (NICE); NIHR Health Technology Assessment (NIHR HTA) programme websites; Campbell Collaboration website; Evidence for Policy and Practice Information (EPPI) Centre website and it’s database of systematic and non-systematic reviews of public health interventions (DoPHER).
the specialty-specific systematic reviews and, therefore, allow normative conclusions on the effectiveness of identifying indicators.

Details from key previous literature reviews of quality and safety indicators were collated (Appendix B, p297) with a summary of these previous reviews given in Table 12.

**Table 12. Selected details from sample of previous review of indicators**

<table>
<thead>
<tr>
<th>Review</th>
<th>Date of publication</th>
<th>Search terms</th>
<th>Resources</th>
<th>Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>International Quality Improvement Programme (IQIP)</td>
<td>Since 1985</td>
<td>Not stated</td>
<td>Not stated</td>
<td>Nearly 700 indicators covering four care settings</td>
</tr>
<tr>
<td>HCUP QIs</td>
<td>1994/1998</td>
<td>Not stated</td>
<td>Not stated</td>
<td>Final list of 33 quality measures</td>
</tr>
<tr>
<td>National Indicators of Safety and Quality Project (Australia)</td>
<td>ACSQHC established in 2000</td>
<td>Not stated</td>
<td>Not stated</td>
<td>152 papers on health</td>
</tr>
<tr>
<td>AHRQ QIs</td>
<td>2001</td>
<td>MeSH terms “hospital, statistics, and methods” and “quality indicators”</td>
<td>Embase, Medline, hand search, author search</td>
<td>181 articles on potential indicators, with 27 explicitly defining novel indicator</td>
</tr>
<tr>
<td>Millar et al (OECD HCQI project)</td>
<td>2004</td>
<td>n/a (reviewed 7 preview indicator sets)</td>
<td>Based on previous indicator searches</td>
<td>21 indicators selected</td>
</tr>
<tr>
<td>SimPatIE</td>
<td>2007</td>
<td>Patient safety, indicator, risk, harm and test*, Usage/Use, Apply, Valid</td>
<td>Pubmed and google scholar</td>
<td>42 potential indicators with 24 recommended for all or part of Europe</td>
</tr>
<tr>
<td>CPSSQ (Imperial College, London)</td>
<td>2008</td>
<td>Two dimensional search string pertaining to [patient safety indicators] and [adverse events]</td>
<td>Included: Medline, Embase, grey literature, government websites</td>
<td>91 articles on patient safety indicators</td>
</tr>
<tr>
<td>Joint Commission’s NHQM</td>
<td>2009 (v2.6b) based on work from 1999</td>
<td>Not stated</td>
<td>Not stated</td>
<td>ORYX initiative has over 8,000 measures. CMS chose 21 measures from shortlist of 39</td>
</tr>
</tbody>
</table>

Notes: NHQM: National Hospital Quality Measures; ACQHC: Australian Council for Safety and Quality in Health Care

### 3.4.4. Literature search resources

#### 3.4.4.1. Electronic databases

Electronic databases were searched to identify both peer-reviewed journal articles and other research such as conference proceedings and unpublished work. There were two main processes regarding the use of electronic databases:

- An extensive systematic search of Institute of Scientific Information (ISI) Web of Science, Medical Literature Analysis and Retrieval System Online (Medline) and Excerpta Medica Database (Embase) databases – the latter two through the OVID interface – from inception to the date of the search.
- Searching databases of previously compiled indicators such as: AHRQ’s Clearinghouse (which included measures used by, for instance, the Canadian Institute for Health Information, Australian Council on Healthcare Standards, British Medical Association);\textsuperscript{158} OECD’s Health Care Quality Indicators (HCQI) project;\textsuperscript{130} the EU sponsored SIMPATIE project and EUPHORIC measures;\textsuperscript{159} HCUP refined indicators;\textsuperscript{80} and Queensland Health: Clinical Practice Improvement Centre measures.\textsuperscript{160}

### 3.4.4.2. Other data sources

To complement the searches of electronic databases, I also used the following methods:

- An online search, including a selection of government and patient safety organisation websites (Appendix C), to identify grey literature, in the form of research and technical papers and reports, conference abstracts, government reports and committee working documents were obtained from this search.

- Searching bibliographies of key articles to identify additional papers which met the eligibility criteria.

The Physiotherapy Evidence Database was not searched as previous research found it did not include new citations relating to these indicators.\textsuperscript{161}

### 3.5. Selecting indicators

#### 3.5.1. Considerations on indicator selection

##### 3.5.1.1. Previous indicator selection

AHRQ had a criteria that “in order to be retained as a potential PSI, at least one of the first three studies needed to demonstrate a positive predictive value of at least 75%, meaning that 3 out of 4 patients identified by the measure did indeed have the complication of interest”.\textsuperscript{40} However the lack of similar studies in the UK on PSIs let alone other specialty-specific indicators meant that such a strong criterion was not possible. As a looser criterion, AHRQ also stated that indicators should have been used effectively in the past, and/or have high potential for working well with other indicators currently in use. In the updating of the initial indicators, the HCUP II QI project has a criterion that at least 1% of hospitalised patients and/or 20% of providers are captured.\textsuperscript{74}
**3.5.1.2. Project protocol for indicator selection**

The intentional differences in the chosen specialties dictated a different set of criteria for selecting the indicators and any specific criteria are presented in the relevant chapters. However, generic criteria based on the findings from previous literature discussed earlier are that indicators should:

a. constitute a bundle, defined here as at least 6 indicators;

b. include process and outcome indicators;

c. focus on an area with known variation in performance;

d. cover adverse events affecting at least 1% of hospitalised patients and/or 20% of providers.

Alongside the author, an expert in administrative data, a public health doctor and a senior doctor in the relevant specialty fed into the development of the indicator set. I acknowledge that the resulting set of indicators would not necessarily be unique; however, this does not affect the ability to determine the feasibility of applying such indicators. I have demonstrated how these are also aligned with the analytical framework in (Table 13), although the final three domains on applicability (‘Data availability’, ‘Reporting burden’, ‘External and ecological validity’) are not included despite these inherently being partially addressed by the characteristics of indicator bundles offering a wider range of information and yet being based on the same data source.
Table 13. Use of multiple measures and mix of process and outcome measures to address analytical framework

<table>
<thead>
<tr>
<th>Domain</th>
<th>a. Multiple measures</th>
<th>b. Process and outcomes</th>
<th>c. Known variation</th>
<th>d. Prevalent</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Impact</td>
<td>More comprehensive understanding of impact</td>
<td>Understanding of both impact on hospitals (process) and patients (outcome)</td>
<td>-</td>
<td>Evidence of high impact</td>
</tr>
<tr>
<td>2. Amenable</td>
<td>-</td>
<td>Associations between processes and outcomes suggest amenability</td>
<td>Suggests possibility to improve to peer performance</td>
<td>-</td>
</tr>
<tr>
<td>3. Face and content validity</td>
<td>Cover more facets of subject of evaluation</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>4. Construct validity</td>
<td>Can test for associations between measures</td>
<td>Can test for associations between process and outcome measures</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>5. Precision</td>
<td>-</td>
<td>-</td>
<td>Ability to identify differences</td>
<td>-</td>
</tr>
<tr>
<td>6. Minimum bias</td>
<td>-</td>
<td>Process measures less likely to be affected by case mix</td>
<td>-</td>
<td>Higher numbers reduce effect of random variation</td>
</tr>
</tbody>
</table>

Notes: Sources and Notes in Table 8, p67.

3.5.2. Defining indicators

An established requirement is that the indicator is based on agreed definitions, and described exhaustively and exclusively. The indicators used in this study are, therefore, set out in detail with key assumptions explained.

3.6. Applying indicators

In this section I describe the statistical methods common to the application of indicators to both specialties. The alignment of these statistical tests with the analytical framework is set out in Table 14, although again the final three domains on applicability (‘Data availability’, ‘Reporting burden’, ‘External and ecological validity’) are not included since these are addressed through the application itself rather than the results.
### Table 14. Statistical tests to address the analytical framework

<table>
<thead>
<tr>
<th>Domain</th>
<th>Statistical technique</th>
<th>Output</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Impact</td>
<td>Descriptive analysis</td>
<td>National rates of adverse events</td>
</tr>
<tr>
<td>2. Amenable</td>
<td>Regression (performance v. structural factors)</td>
<td>Effect size and significance of correlation</td>
</tr>
<tr>
<td>3. Face and content validity</td>
<td>n/a</td>
<td>n/a</td>
</tr>
<tr>
<td>4. Precision</td>
<td>Descriptive analysis</td>
<td>Range of scores, by group</td>
</tr>
<tr>
<td></td>
<td>Statistical testing</td>
<td>Mean number of cases, per unit</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Presentation on funnel plots</td>
</tr>
<tr>
<td>5. Construct validity</td>
<td>Regression</td>
<td>Spearman rank correlation, rho, and significance</td>
</tr>
<tr>
<td>6. Minimum bias</td>
<td>Multi-level regression</td>
<td>Risk-adjusted rates, ( r^2 ) (variation explained by model) and ( c )-statistic (discriminatory power at admission level)(^{162})</td>
</tr>
<tr>
<td></td>
<td>Comparison of risk-adjusted and crude rates</td>
<td>Spearman rank correlation, rho, and significance; % change in outlier performance(^{163})</td>
</tr>
<tr>
<td></td>
<td>Descriptive analysis</td>
<td>Stratification of key results</td>
</tr>
<tr>
<td></td>
<td>Regression analysis including hospital-level coding variables</td>
<td>Coding- and risk-adjusted rates</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Spearman rank correlation, rho; and significance</td>
</tr>
<tr>
<td></td>
<td></td>
<td>% change in outlier performance</td>
</tr>
</tbody>
</table>

Notes: Sources and Notes in Table 8, p67.

### 3.6.1. Descriptive statistics

For most of the analyses, the first step is to produce crude statistics, including bivariate associations. The methods used are summarised in Table 15.

### Table 15. Simple statistical tests

<table>
<thead>
<tr>
<th>Test</th>
<th>Data type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pearson chi-squared, ( X^2 )</td>
<td>Categorical</td>
</tr>
<tr>
<td>Mann Whitney U test</td>
<td>non-normally distributed ordinal</td>
</tr>
<tr>
<td>t-tests (ANOVA)</td>
<td>normally distributed continuous</td>
</tr>
<tr>
<td>Pearson correlation, ( r )</td>
<td>non-normally distributed continuous</td>
</tr>
<tr>
<td>Spearman rank correlation, ( r )</td>
<td>non-normally distributed ordinal</td>
</tr>
</tbody>
</table>
3.6.2. Regression models

Regression analyses were used to calculate odds ratios (ORs) for both unadjusted and case-mix adjusted indicators. The indicators in this project are dichotomous in nature (with each admission either having or not having the complication in question) and, as such modelling is based on the binomial distribution. Each relevant hospital admission is flagged with a “0” or “1”, referring to the absence or not of the quality or safety events. The binomial distribution has mean \(n \cdot p\) and standard deviation \(\sqrt{n \cdot p \cdot (1-p)}\), where \(p\) is the proportion of cases and \(n\) the sample size.

The logistic regression can be used to predict the odds at an admission-level (Formula 1) and odds ratio (OR), which is defined as the odds from one group of admissions divided the odds from another group (Formula 2). Odds and ORs have a floor of 0 and no ceiling.

\[
\text{Odds}_i = \frac{\text{Probability (admission i has an event)}}{\text{Probability (admission i does not have event)}} \quad \text{Formula 1}
\]

\[
\text{OR}_{jk} = \frac{\sum_{i \in j} \text{Odds}_i}{\sum_{i \in k} \text{Odds}_i} \quad \text{Formula 2}
\]

Where:

\(i = \text{admission within a wider set, say, } A; j,k = \text{groups of admissions within set } A;\)

Regression analyses are used to predict the presence of an event based on predictor variables, which can be either continuous or categorical data. The expected value of the response variable is fit to these predictor variables, e.g. the probability of death is estimated for different patient age groups. Using natural logarithm transformation (and inverted back using the exponential function), a logistic regression can predict probability of an event as a continuous variable even though the observed events are categorical (binary). The PROC LOGISTIC function is used within the SAS software.

3.6.3. Generalised linear models

When there is evidence that the outcome is affected at different levels – in this case, both patient-level and provider-level, a multi-level model is required. Where this was the case, I fitted generalised linear mixed models using SAS’s PROC GLIMMIX function. The clustering of patients may affect results; however, where I report provider-level variation with no hospital level explanatory factor I do not adjust for clustering (in the reported model) since the effects were found to be small.
PART I: Introduction

3.6.4. Unadjusted indicator rates

In the first instance, crude rates (at national and provider levels) were calculated for each indicator. These simple proportion or crude rates (CR) are given by Formula 3 with the 100(1-α)% confidence interval limits given by Formulae 4a and 4b.\(^{164}\)

\[
CR_i = \frac{O_i}{n_i} \quad \text{Formula 3}
\]

\[
CR_{\text{lower}} = \frac{O_{\text{lower}}}{n} = \frac{O_i(1 - \frac{1}{90})^\frac{z}{\sqrt{n_i}^3}}{n} \quad \text{Formula 4a}
\]

\[
CR_{\text{upper}} = \frac{O_{\text{upper}}}{n} = \frac{O_i(1 - \frac{1}{90(0+1)})^\frac{z}{\sqrt{n_i}^3}}{n} \quad \text{Formula 4b}
\]

Where:

\(i = \text{hospital 1,…;}\)   \(n_i = \text{total number of admissions at hospital } i\)

\(z = 1 – \alpha/2\) percentile of a standard normal distribution

\(O_i = \text{observed number of admission at hospital } i\) meeting the numerator criteria (e.g. receiving scan)

3.6.5. Risk adjustment

Previously cited risk factors were collated in the systematic reviews. A priori variables were retained unless they threatened model convergence. This allowed consistency in the case-mix adjustment across the different measures. In other words, all candidate variables were retained, even if not statistically significant, as the goal was confounder control.\(^{162}\)

3.6.6. Goodness of fit

A common measure of goodness of fit is that described by Hosmer and Lemeshow; however, recent analyses have suggested this test is likely to produce significant results with large data sets and it is unclear how important the imperfect calibration is for a typical hospital in practice.\(^{165}\) I inspected standard errors for non-convergence, co-linearity, and over-fitting, including only main effects. As reported in the analytical framework above, I reported the c- and \(r^2\)-statistics for each model:

- the c-statistic measures how well models discriminate between cases with an adverse event to those without (0.5 indicating no ability to discriminate and 1 for perfect discrimination);
- the \(r^2\)-statistic is a measure of how much of the variation is explained by the model (0 indicating none of the variation being explained and 1 indicating all explained).\(^{166}\)
3.6.7. Standardisation of indicator results

A standardised ratio is the observed number of events relative to the number of events that would be expected if predictive chance of events was applied to the particular observed population characteristics. A common example is the standardised mortality ratio (SMR). The indirectly standardised ratio (ISR) is given in Formula 5, with confidence intervals given in Formulae 6a and 6b. From these standardised rates, an indirectly adjusted rate (IAR) can be calculated by applying the ISR to the population crude rate (Formula 7).

\[
IR = \frac{O}{E} = \frac{\sum_p O_p}{\sum_p E_p} = \frac{\sum_p O_p}{\sum_p n_p \lambda_p}
\]  
Formula 5

\[
ISR_{lower} = \frac{O_{lower}}{E} \quad \text{Formula 6a} \quad ISR_{upper} = \frac{O_{upper}}{E} \quad \text{Formula 6b}
\]

\[
IAR = ISR \cdot CR_{population}
\]  
Formula 7

where:

- \(O_p\) is the observed number of events in the subject population in group \(p\);
- \(E_p\) is the expected number of events in the subject population in group \(p\) given the standard rates;
- \(n_p\) is the number of individuals in the subject population in group \(p\);
- \(\lambda_p\) is the group-specific rate estimated rate for group \(p\).

\(O_{upper}\) and \(O_{lower}\) are defined above in Formulae 4a and 4b.

This method assumes both: (1) a homogeneity within each patient group; and (2) that, overall, the expected number of events is that of the observed level. The group-specific rates \(\lambda_i\) can be calculated either directly (subject population rates applied to a standard population) or indirectly (applying standard or calculated rates to the structure of the subject population). This project uses the latter, due to the advantage of being more robust when there small number of events of interest and due to a lack of standard/reference population making the former unfeasible.

3.6.8. Funnel plots

I used funnel plots, with 95% and 99.8% confidence levels. The main advantage of this graphical representation is that it can reflect the fact the confidence intervals for different sizes of providers.
3.7. Setting

England was chosen to evaluate the feasibility of applying specialty-specific indicators, which complied with the inclusion criteria that the setting was outside the USA and based on administrative data which used the ICD-10 diagnosis coding framework. As well introducing the setting, I show below that this choice also complies with the analytical framework criteria of high impact and being amenable.

3.7.1. Structure of English NHS

In England, 60 years after its creation, the National Health Service (NHS) provides 64 million outpatient and A&E attendances and, in general practice, 290 million consultations per year. This care is provided by the 1.35 million staff who work for the range of providers including NHS hospitals (NHS trusts or NHS foundation trusts), GPs, dentists, and private sector and voluntary sector organisations. Following the reforms of 2012, the policy direction is set by the Department of Health (DH), with leadership for the commissioning system assumed through the independent organisation NHS England.

3.7.2. Quality and safety in England

3.7.2.1. Prevalence of quality and safety in England

Recognition of the patient safety issue in England came with the publication of a seminal report in 2000. Five years later, a study by the National Audit Office reported that the cost to the NHS of patient safety events included: £2 billion a year in extra bed days; an additional £1 billion to this expenditure due to hospital acquired infections; and £423 million in settled clinical negligence claims (National Audit Office 2005).

3.7.2.2. Quality and safety regulation regime in England

In England, the Care Quality Commission (CQC) has acted as the independent quality regulator for all health and social care services in England since 2009. Previously, this role was conducted by the Healthcare Commission, the Commission for Social Care Inspection and the Mental Health Act Commission.

A system for reporting patient safety incident reports – the National Reporting and Learning System (NRLS) – was established in 2003, with the intention that the information is used to develop tools and guidance to help local improvements in safety. More than four million incident reports have

\[d\] The NRLS also cover NHS trusts in Wales
now been submitted by healthcare staff. In 2010, it became mandatory for hospital trust to report serious incidents to the Care Quality Commission.171

3.7.2.3. Quality and safety monitoring in England
The monitoring of quality and safety in England is developing; however, few projects have used administrative data. The Government introduced the NHS Outcomes Framework in 2012 (with full role from 2013) to provide an overview of performance and act as a catalyst for driving quality improvement and outcome measurement.172 The Framework replaced the old regimes such as ‘vital signs’ and ‘national indicators’. The NHS Executive introduced a set of clinical indicators based on administrative data in 1999.173 Yet a systematic review of published articles on patient safety indicators published in 2008 identified only two studies23,174 describing adverse event measurement and monitoring in British hospitals.38

3.7.3. Focus on acute care

3.7.3.1. Definition of acute setting
‘Primary care’ services are typically the first contract point within the NHS and provided by, for example, GPs, NHS-walk-in-centres, dentists, and pharmacists. Secondary, or ‘acute’ care, is provided by hospitals.

3.7.3.2. Framework for identifying the setting
Administrative data systems exist in many healthcare settings, such as family (general) practice and hospitals, and can be used to analyse the quality and safety of the setting that provides the data or other settings. For instance, hospital administrative data can evidently be used to measure hospital performance, but by analysing emergency admission rates, for example, inferences about the quality of community care could be made. The following paragraphs set out why the project focuses on hospital data and hospital performance; although I do not argue that other data resources and settings are less important. To show the value in this scope, I again recall the criteria above on ‘importance’ regarding the room for improvement and being amenable.

3.7.3.3. Importance for hospitals
As discussed above (paragraph 1.3.2, p28), the importance of quality and safety for hospitals is well established.

3.7.3.4. Amenable to indicator-use
The use of quality and safety indicators based on administrative data is relatively recent, especially outside the USA, and so little work has gone into evaluating whether they are an effective tool to
improve care. However, there have been some promising examples, such as one hospital that detected poor performance on the HCUP indicators and were able to amend their procedures to improve appropriate utilisation. There is also evidence that similar improvement programmes – such as using other computerised clinical information – can be well-received by hospital providers, improve error detection and could reduce the incidence of adverse events in hospitals.

3.7.3.5. Scope to improve use of indicators in this setting
Importantly, there also remains scope to improve the use of indicators based on administrative data in this setting. A systematic review of patient safety indicators in 2008 identified 165 publications relating to patient safety in hospitals, of which only 51 were original research. Only one cross-sectional study (a survey for determining hospital staff’s attitude to safety culture) and on case control study (using administrative data to identify potential quality problems) investigated validation.

3.7.3.6. Findings on quality and safety in English hospitals
The existence of significant variations in quality and safety of care in English hospitals has been well established. In particular, annual reports by Dr Foster Intelligence – a joint venture with the Department of Health – have suggested wide-spread variation over a range of conditions and measures.

3.8. Data
3.8.1. Introduction to Hospital Episode Statistics
The Hospital Episode Statistics captured almost 18 million episodes of admitted patient care in 2012-13, with each record covering the continuous period during which the patient is under the care of one consultant (Finished Consultant Episode, FCE). These FCEs are linked together into hospital spells which can be further linked to any spells resulting from the transfer of a patient to another NHS hospital to form ‘superspells’ (referred to here as admissions). Diagnoses are recorded using the International Statistical Classification of Diseases and Related Health Problems, tenth version (ICD-10) and procedures are coded using the Office of Population Censuses and Survey’s Classification of Surgical Operations and Procedures, fourth version (OPCS-4).

This data constitutes Hospital Episode Statistics (HES) which were conceived in 1987, to replace a system based on collecting data from around 10% of patients. The data are cleaned according to published rules.
3.8.2. Ethical considerations
The Dr Foster Unit at Imperial has permission to hold HES data under Section 251 (formerly Section 60) granted by the National Information Governance Board for Health and Social Care (NIGB, formerly the Patient Information Advisory Group). The Unit also have approval for using these data for research from the South East Research Ethics Committee.

3.8.3. Software
The following software programmes were in the project: Endnote X3, X5 and X6 (Endnote Carlsbad, CA); Microsoft Office Word and Excel 2007 and 2010; and SAS Version 9.2 TS Level 2MO.
Part Two:

Stroke care
## Chapter 4. Stroke literature review

|---------------------|---------------------------------|--------|-------------------------------|------------|--------|-----------------------------------------|------------|--------|-------------------------------------|------------|-----------|------------|

---

87
PART II: Stroke Care

Overview

Context
The previous part of the monograph outlined the rationale and analytical approaches for using routinely-collected administrative data, focused at the specialty-level, to evaluate the quality and safety of hospital care. To understand the feasibility of meaningfully applying the indicators to a particular specialty (acute stroke care), a consolidated understanding of previous measurement efforts is required. This objective was met following the systematic review methodology described in this chapter.

Methods
A systematic review of original studies that applied, evaluated or validated stroke-specific indicators based on administrative data. Indicator and study details were collated.

Findings
The initial search identified 1,999 unique citations and I identified 99 studies, and 20 broad groups of indicators. Most measures had been applied in the USA (n=64) and Canada (n=11) with few studies describing any substantial validation of the indicators, and the application of the indicators – in terms of definitions of stroke and risk-adjustment models – inconsistent. The use of stroke-specific indicators is patchy, in terms of the countries that have applied them and the indicators used. There is potential to make more meaningful inferences about the quality and safety of acute stroke care by applying a more comprehensive set of indicators and being more explicit about the assumptions used.

The literature review followed a similar process to a review of all patient safety measure based on routinely-collected hospital data which identified only a similar magnitude of studies (n = 124) despite its far broader scope.

What this chapter adds

- The first comprehensive review of stroke-specific indicators based on administrative data.
- The first review of how the stroke indicators have been applied, including strategies for case ascertainment, risk-adjustment and validation.
4.1. Background on stroke care

4.1.1. About stroke
Strokes are caused by an interruption of the blood supply to the brain, either through a clot (ischaemic) or bleed (haemorrhagic). The subsequent lack of oxygen and nutrients can cause damage to the brain tissue, commonly resulting in sudden weakness or numbness of the face, arm or leg. The outcome from stroke will, along with patient characteristics, depend on the area of the brain affected and the severity of the interruption.\textsuperscript{182} An indicative stroke care pathway is included as Figure 5.

4.1.2. Importance

4.1.2.1. Burden of stroke
Stroke is the second leading cause of death worldwide, accounting for an estimated 5.7 million fatalities annually, equivalent to 1 in 10 deaths.\textsuperscript{182}

4.1.2.2. Treatment of stroke
Specialist training for stroke care is a recent phenomenon.\textsuperscript{183} The development of the area as a specialty has coincided with dramatic changes in the management of stroke care due to burgeoning evidence of effectiveness of certain treatments. For instance, a systematic review in 1997 indicated that organisation of inpatient care in designated wards (stroke units), compared to conventional care, reduced mortality and disability despite no routinely employed medical or surgical interventions at that point.\textsuperscript{184,185} However, in the subsequent years such interventions were developed and, for instance, by 2007 the use of thrombolysis “clot busting” treatment for certain stroke types (ischaemic) had been approved in England.\textsuperscript{186} This, in turn, changed the importance of stroke patients receiving urgent scans so that their stroke type could be determined within the time window for administering thrombolysis.\textsuperscript{151}

4.1.2.3. Amenable to healthcare
Previous research has suggested that stroke outcomes can be improved through the delivery of improved healthcare. Perhaps as a result of the developing nature of stroke care with lag time in some organisations and settings, unacceptable variations in the organisation and delivery of care exists.\textsuperscript{187} For instance, a national (English) report also recently highlighted both the enduring variations in hospital care and that some levels of care remain unacceptably low.\textsuperscript{151} This phenomenon is not restricted to England.\textsuperscript{188}
Figure 5. Indicative recommended stroke care pathway

**Urgent response:** First contact (e.g. GP or emergency call). The patient goes as soon as possible to specialist stroke unit (either directly or via A&E).

**Hyper-acute treatment:** Brain scan as soon as possible to determine the type of stroke and whether patient is eligible for, e.g., thrombolysis treatment.

**Acute care:** The patient should spend the rest of their time in hospital in specialist stroke unit – where a multi-disciplinary team work with them to support recovery (e.g. doctor, nurses, physiotherapist, occupational therapist, dietician, social worker).

**Post-hospital care:** On leaving hospital – rehabilitation may be continued in the community, and further support may be necessary (housing, adaptations to the home, personal care etc.) Patients should be reviewed at 6 weeks, 6 months and annually thereafter.

< 4 hours

- **Prioritised ambulance response**
- **Brain scan**
- **Thrombolysis (if needed)**

Urgent care pathway for strokes identified early

- **Intensive monitoring on stroke unit**
- **Intensive rehabilitation**
- **Community based specialist rehabilitation**
- **Home with on-going rehabilitation package, including assessment and social care**

Source: Adapted from National Audit Office (2010)
4.2. Introduction

4.2.1. Stroke as a specialty
As a specialty, ‘Stroke’ is a subset of internal medicine dealing with the treatment of strokes. Stroke is a new specialty with attending physicians previously – and often still – being geriatric or neurology specialists.

4.2.2. Stroke measurement

4.2.2.1. Demand for indicators
There has been widespread demand for better measurement of the burden of stroke and the quality of care administered. The WHO have actively supported countries to understand their stroke mortality rates whilst also highlighting a general lack of quality and safety indicators, particularly in developing countries. At a national-level, taking the example of England, the DH have introduced stroke-specific measures of discharge home, survival rates and emergency readmissions to hospitals within 28 days of discharge for stroke. The Department also reiterated that, compared with other countries, there were relatively poor outcomes in England across some measures of stroke and, as such, the area was earmarked as a priority. Academics have also called for indicators to measure stroke care performance.

4.2.2.2. Existing measures
To date, the most comprehensive and high-profile study to benchmark stroke care have been based on specific stroke registers and prospective studies. National stroke audits are being used across the world, such as in England, Argentina, and Canada. More detail on measurement in the former of these nations (England), is given in the following chapter (paragraph 5.1.1.2, p117). In late 2009, a small indicator set was introduced for hospitals to monitor and review which included in-hospital stroke mortality.

Previous research has suggested that the use of quality indicators in stroke care could improve performance. For instance, improvements have been attributed to the introduction of a registry in the USA with 10 performance measures, another USA-based registry, and the NSSA. Despite the promising outcomes from these performance measurement activities, little use has been made of the vast amounts of detailed, patient-level, administrative data that is routinely collected in many nations to measure the quality and safety of individual specialities such as stroke care. As such, there would be value in documenting what stroke-specific indicators have been used to date and how these have been applied.
PART II: Stroke Care

4.3. Objectives

To produce a comprehensive list of indicators of the quality and safety of hospital stroke care that can be based on administrative hospital data, and provide a critical assessment of how these have been applied previously. This latter objective is designed to inform the quantitative analyses presented in the following three chapters.

4.4. Methods

4.4.1. Search resources

An outline of the systematic methodology was given in a previous chapter (paragraphs 3.4.2-3.4.3, p72-). In summary, the approach used in this review is developed from the methodology of Tsang and colleagues on reviewing patient safety measures based on routinely collected hospital data.\textsuperscript{67}

The method used to search for peer-reviewed journal articles and other research is described earlier (paragraph 3.4.4, p74). An initial search of the databases – designed to identify peer-reviewed journal articles – was performed between May and July 2010 and also later in the review process in May 2011.

4.4.2. Validation of search results

As a further check of completeness, I also reviewed the extracted indicators against measures in an American list of approved stroke and stroke rehab measures, even though they were not designed specifically for administrative data.\textsuperscript{200} Where a study was based on a database similar to administrative data – such as a Cleveland-based study\textsuperscript{201} which included five measures that could all be applied to administrative data – these measures were cross-referenced against the compiled list of potential indicators to ensure they were already included.

4.4.3. Definitions

4.4.3.1. Definition for administrative data

Recall from paragraph 2.1.1.1 that ‘administrative data’ is defined, for the purpose of this review and to meet the ambition of the study, using the inclusion criteria: routinely collected data covering more than one hospital; used for administrative purposes (e.g. for reimbursement); and can be replicated using the internationally-recognised ICD-9 or ICD-10 diagnoses coding framework.

As a principle, where a study was based on administrative data linked to another supplementary source, the individual measures described in the paper were reviewed to assess whether they could be replicated with administrative data alone. This ensured that no potentially important indicator
was excluded and that validation studies, where results from administrative data are compared to an augmented or different dataset, were included.

4.4.3.2. Definition of quality and safety
As discussed earlier (paragraph 1.1.4), providing an exact definition of quality and safety is difficult and its application for the purpose of this review on stroke care is possibly best defined by the treatment of contentious indicators:

- Length of stay is included as, while in some instances this is considered a measure of hospital efficiency rather than treatment effectiveness, some variations – such as a dichotomous indicator of long length of stay – are more likely to be considered a proxy for aspects of quality. One study, which described length of stay as an economic measure, in fact found that it was positively correlated with quality measures.202

- Hospital charges are not considered as a quality indicator since this is a compound measure of patient characteristics, length of stay, and treatments provided – which can be picked up as separate indicators and risk adjustments – and is a more direct measure of hospital efficiency.

- Clinically associated complications of stroke that are not amenable to healthcare – such as a study on fractures following stroke – are not included.203

4.4.4. Search strategy

4.4.4.1. Electronic databases
The following three electronic databases were interrogated: Ovid MEDLINE(R) In-Process & Other Non-Indexed Citations and Ovid MEDLINE(R) 1946 to Present; Ovid Embase Classic+Embase 1947 to 2011 May 15; and Web of Science (to present).

4.4.4.2. Search terms
The search terms used were informed by previous stroke literature reviews.204,205 The databases were interrogated using a two-dimensional search string using the ‘AND’ conjunction to identify papers relating to stroke which mention either a quality indicator or routine data (Figure 6). Medline and Embase were interrogated using both indexing terms (MESH and EMTREE, as shown in the Figure) and also using a two-dimensional search string akin to that used for Web of Science to check if the use of the index terms were systematically omitting results.
Terms for TIAs were also included to accommodate studies which were focused on these ‘minor strokes’ but may include stroke-specific indicators. As with similar reviews, only English language articles were included, and non-human studies were removed. 57,161

Figure 6. Search terms

<table>
<thead>
<tr>
<th>Web of Science</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Indicator* OR &quot;administrative data&quot; OR &quot;routine data&quot; OR &quot;routinely collected data&quot; OR &quot;hospital episode statistics&quot; OR &quot;hospital episodes statistics&quot; OR &quot;claims data&quot; OR &quot;International Classification of Diseases&quot; OR &quot;International Statistical Classification of Diseases&quot; OR ICD NOT &quot;implantable cardioverter&quot; AND Stroke OR Cerebro<em>vascular OR CVA OR Apoplexy OR (brain AND vascular) OR (transient isch</em>mic attack).</td>
<td></td>
</tr>
<tr>
<td>Embase</td>
<td></td>
</tr>
<tr>
<td>‘health care quality’ OR ‘clinical indicator’ OR ‘performance measurement system’ OR ‘international classification of diseases’) AND ‘stroke’ OR ‘transient ischemic attack’.</td>
<td></td>
</tr>
<tr>
<td>Medline</td>
<td></td>
</tr>
<tr>
<td>‘quality indicators, health care’ or ‘international classification of disease’ AND ‘stroke’ OR ‘ischemic attack, transient’.</td>
<td></td>
</tr>
</tbody>
</table>

4.4.4.1. Duplicates

Duplicate citations were identified using Endnote reference management software by identifying entries with identical authors, date, journal, and title.

4.5. Review process

4.5.1. First stage review criteria

4.5.1.1. Inclusion and criteria for review of indicator use

Recall that literature on quality and safety measures were retained if (1) based on routinely collected hospital administrative data; and (2) applied, evaluated or validated a potential or actual indicator of the quality or safety of hospital stroke care. To meet this latter criterion, an included study had to provide numerical results on performance against the indicator.

Specifically, as a first stage, abstracts were reviewed and candidate papers were excluded if:

× None of the described indicators related to an acute stroke so, for instance, excluding measures relating to surgical procedures for prevention of stroke;
None of the described indicators explicitly relate to the quality and safety of health care provided, such as measures of cost (and also stroke incidence given this is a wider measure of public health). As a result, readmission for stroke is included but neither stroke incidence nor those complications of stroke not associated with quality (such as Generalised Convulsive Status Epilepticus, GCSE) are included; and

The study was not based on routinely-collected hospital administrative data (as defined above).

Where there was insufficient detail in the abstracts to ascertain whether the criteria above were met, the study was included in the second stage (full text) review.

### 4.5.1.1. Inclusion of additional papers describing indicators

Secondly, I separately collated studies which described indicators that could be applied to administrative data but no actual database was used (and so failing the first stage criteria to provide numerical results) to ensure a well-defined comprehensive list of indicators.

### 4.5.2. Second stage review criteria

For the second stage, the remaining articles had the same exclusion criteria applied to the full texts. This stage provided a form of quality assurance by double-reviewing many of the studies whilst also giving the opportunity to collate a range of additional information from the papers. As well as the details of the publication (such as author, title, year of publication) a standard template was used to extract information relating to the studies (Table 16). Information on validation techniques applied and any key limitations were also recorded where available. The heterogeneity of the studies being considered meant that no meta-analysis would be possible so the results were not systematically recorded.
### Part II: Stroke Care

#### Table 16. Information extracted

<table>
<thead>
<tr>
<th>Study section</th>
<th>Data field</th>
</tr>
</thead>
<tbody>
<tr>
<td>Design</td>
<td>Country</td>
</tr>
<tr>
<td></td>
<td>Details of cohort</td>
</tr>
<tr>
<td></td>
<td>Longitudinal</td>
</tr>
<tr>
<td></td>
<td>Key Limitation(s)</td>
</tr>
<tr>
<td>Data</td>
<td>Database(s)</td>
</tr>
<tr>
<td></td>
<td>Year(s) of data</td>
</tr>
<tr>
<td></td>
<td>Diagnosis coding framework</td>
</tr>
<tr>
<td>Variables</td>
<td>Quality and safety indicator(s)</td>
</tr>
<tr>
<td></td>
<td>Key covariate(s)</td>
</tr>
<tr>
<td>Case ascertainment</td>
<td>Diagnosis field(s)</td>
</tr>
<tr>
<td></td>
<td>Diagnosis codes</td>
</tr>
<tr>
<td>Statistical methods</td>
<td>Risk adjustment</td>
</tr>
<tr>
<td></td>
<td>Validation technique(s)</td>
</tr>
</tbody>
</table>

#### 4.5.3. Data extraction and review

**4.5.3.1. Analytical tools**

I catalogued and reviewed the extracted details using Endnote versions X3 and X4 (EndNote Carlsbad, CA) reference management software. A standard form was used to collate the information for the second stage (review of full texts). Similarly, a standard template was used to record the different categories of measures used. Statistical analyses of study details, such as trends in number of publications, were performed using Microsoft Office Excel 2010 (Microsoft Corp., USA).

**4.5.3.2. Comparative analyses**

In discussing the descriptive statistics on the literature, I use comparisons to previous literature reviews on patient safety indicators, predominantly that of Tsang et al,\(^66,67\) to provide a benchmark to judge whether the specialty-specific nature of the review, and the specialty in question (stroke), have revealed atypical results. A comparison to the obstetrics literature review (Chapter 7) is included in the discussion (Chapter 11).

The information from the studies was compared as a whole and in subsets since certain purposes, for instance, would require specific risk-adjustment models and strategies for identifying stroke episodes. For example, where the purpose of the study is to validate the identification of safety issues against medical notes, it is unlikely that you would want to adjust for case-mix, whereas for comparisons between providers such a risk-adjustment would be important.
4.6. Results

4.6.1. Identification of studies
The initial search identified 1,999 articles of which 206 duplicates were removed. Of the remaining 1,793 articles screened, the full text was retrieved for 169 studies with the most common reason for exclusion being: non-relevant (n=1243); the study describes an indicator not applied to routinely-collected administrative data (n=140); evaluated only the prevalence of stroke (n=78); and investigated only the accuracy of diagnosis coding of stroke (n = 33).

During the second review stage, a further 88 studies were excluded, although 18 were added from the search of grey literature and review of the bibliographies of key articles, resulting in 99 articles (Figure 7). The information extracted from the articles are summarised in Table 17.
Figure 7. Inclusion and exclusion of articles in the literature review

1,999 records identified through database searching

- Duplicate record?
  - Yes → 206 records excluded
  - No → 1793 abstracts screened

- Eligibility criteria met?
  - No → 1,624 records excluded
  - Yes → 169 full-text articles assessed for eligibility; bibliographies reviewed

18 additional records identified through other sources and bibliographies of key articles

- Eligibility criteria met?
  - No → 88 full-text articles excluded
  - Yes → 99 studies included for extraction of indicator details
Table 17. Characteristics of literature

<table>
<thead>
<tr>
<th>Data years ^</th>
<th>Number of studies</th>
<th>Proportion of all studies</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pre-1995</td>
<td>10</td>
<td>10.1%</td>
<td>Ostbye (1997)\textsuperscript{208}, Shapiro (1994)\textsuperscript{209}</td>
</tr>
<tr>
<td>2000-2004</td>
<td>41</td>
<td>41.4%</td>
<td>Kuwubara (2006)\textsuperscript{212}, Lee (2010)\textsuperscript{213}</td>
</tr>
<tr>
<td>Post-2004</td>
<td>19</td>
<td>19.2%</td>
<td>Kleinendorfer (2009)\textsuperscript{214}, Lichtman (2011)\textsuperscript{215}</td>
</tr>
<tr>
<td>Not described</td>
<td>5</td>
<td>5.1%</td>
<td>Tirschwell (1999)\textsuperscript{216}</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Country of study</th>
<th>Number of studies</th>
<th>Proportion of all studies</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>USA</td>
<td>64</td>
<td>64.6%</td>
<td>Mitchell (1996)\textsuperscript{217}, Borzecki (2010)\textsuperscript{218}</td>
</tr>
<tr>
<td>Canada</td>
<td>11</td>
<td>11.1%</td>
<td>Field (2004)\textsuperscript{219}, Johansen (2006)\textsuperscript{220}</td>
</tr>
<tr>
<td>England</td>
<td>4</td>
<td>4.0%</td>
<td>Laudicella (2008)\textsuperscript{221}, Lazzarino (2011)\textsuperscript{222}</td>
</tr>
<tr>
<td>Taiwan</td>
<td>3</td>
<td>3.0%</td>
<td>Lee (2010)\textsuperscript{213}</td>
</tr>
<tr>
<td>Australia</td>
<td>3</td>
<td>3.0%</td>
<td>Scott (2004)\textsuperscript{223}</td>
</tr>
<tr>
<td>Multi-national</td>
<td>3</td>
<td>3.0%</td>
<td>Mattke (2006)\textsuperscript{130}</td>
</tr>
<tr>
<td>Other single-nation studies</td>
<td>11</td>
<td>11.1%</td>
<td>e.g. Netherlands, Slobbe (2008)\textsuperscript{224}</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Number of primary indicators</th>
<th>Number of studies</th>
<th>Proportion of all studies</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>62</td>
<td>62.6%</td>
<td>Tonarelli (2010)\textsuperscript{225}, Tseng (2009)\textsuperscript{226}</td>
</tr>
<tr>
<td>2 – 3</td>
<td>30</td>
<td>30.3%</td>
<td>Votruba (2006)\textsuperscript{227}, Williams (2004)\textsuperscript{228}</td>
</tr>
<tr>
<td>4 – 5</td>
<td>7</td>
<td>7.1%</td>
<td>Smith (2006)\textsuperscript{229}, Tu (2003)\textsuperscript{230}</td>
</tr>
<tr>
<td>&gt; 5</td>
<td>0</td>
<td>0.0%</td>
<td>-</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Trend analysis</th>
<th>Number of studies</th>
<th>Proportion of all studies</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Longitudinal</td>
<td>30</td>
<td>30.3%</td>
<td>Zhu (2009)\textsuperscript{231}, Volpp (2006)\textsuperscript{232}</td>
</tr>
<tr>
<td>Non-longitudinal</td>
<td>69</td>
<td>69.7%</td>
<td>Zhu (2008)\textsuperscript{233}, Votruba (2006)\textsuperscript{227}</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Diagnosis coding</th>
<th>Number of studies</th>
<th>Proportion of all studies</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>ICD-9</td>
<td>73</td>
<td>73.7%</td>
<td>Adeoye (2011)\textsuperscript{234}, Andaluz (2008)\textsuperscript{235}</td>
</tr>
<tr>
<td>ICD-10</td>
<td>11</td>
<td>11.1%</td>
<td>Cho (2009)\textsuperscript{236}, Gattellari (2009)\textsuperscript{237}</td>
</tr>
<tr>
<td>Other †</td>
<td>15</td>
<td>15.2%</td>
<td>e.g. both ICD-9 and ICD-10, Westert (2002)\textsuperscript{238}</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Stroke diagnosis field</th>
<th>Number of studies</th>
<th>Proportion of all studies</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Primary only</td>
<td>57</td>
<td>57.6%</td>
<td>Gregory (2009)\textsuperscript{239}, Lee (2010)\textsuperscript{213}</td>
</tr>
<tr>
<td>Primary and some secondary</td>
<td>1</td>
<td>1.0%</td>
<td>Meyer (2009)\textsuperscript{240}</td>
</tr>
<tr>
<td>Primary and all secondary codes</td>
<td>6</td>
<td>6.1%</td>
<td>Lanska (1994)\textsuperscript{241}</td>
</tr>
<tr>
<td>DRG-level</td>
<td>3</td>
<td>3.0%</td>
<td>Cleves (1997)\textsuperscript{242}</td>
</tr>
<tr>
<td>Not described</td>
<td>32</td>
<td>32.3%</td>
<td>Censullo (2008)\textsuperscript{243}, Crowley (2009)\textsuperscript{244}</td>
</tr>
</tbody>
</table>

**TOTAL** 99 -

Notes: ^ Where studies covered multiple years of data, the study is categorised according to the most recent year

† Includes studies where coding framework not described and where both ICD-9 and ICD-10 used.
4.6.2. Details from studies

4.6.2.1. Country
Three-in-five published studies were based on data from the USA, with the next most prolific country being Canada (n=11). No other nation had more than 4 studies based on administrative data evaluating this specialty.

4.6.2.2. Year of study
None of the studies were published before 1987; however, this is likely to be an artefact of the commonly used administrative databases not being established before 1988 (paragraph 2.1.1.2, p39). The number of studies has increased over time and around two-thirds (64%) of the studies incorporated data from 2000 or later. Due to time lags in accessing, analysing and publishing studies based on administrative data one might expect a time lag in publishing data and, therefore, the trend was considered before 2007. Using a linear regression of year (of latest data used in each study) against number of studies shows that there has been a substantial and significant increase over time ($\beta = +0.45$, p-value < 0.001) (Figure 8).

Figure 8. Number of publications compared to latest year of data

Note: Year represents the latest date of data used within studies

4.6.3. Study designs

4.6.3.1. Cross-sectional regional analysis
The studies were predominantly cross-sectional in design, with study populations ranging from admissions at a single hospital to multiple national populations and 2 studies presenting the analyses at multiple geographical levels.214,245 At the most aggregated geographical level, there were 3 cross-
sectional descriptive studies comparing performance across countries. Other studies used the regional geographical unit either with the purpose of directly comparing performance, or investigating particular regional-level covariates such as variations in: competition; use of comprehensive regional stroke networks; Health Maintenance Organisation (HMO) penetration; and level of rehabilitation provided by local health authorities within a region.

4.6.3.2. Cross-sectional provider-level analysis

At the provider-level, only two studies were directly focused on highlighting variation between hospitals. Rather, these provider-level studies predominantly focused on investigating the effect of particular hospital characteristics. The most prominent hospital characteristics being studied were:

- teaching status, measured either dichotomously or using the resident-to-bed ratio as a proxy to differentiate between major teaching, minor teaching, and nonteaching hospitals; and

- size, measured through either number of beds or throughput;

- type, including municipal vs voluntary, HMO vs Fee For Service, Veterans Association (VA) vs non-VA, and non-profit vs for profit vs public ownership; and

- level of stroke specialism, which incorporates studies looking at the association in performance with level of stroke accreditation or existence of a specialist stroke unit.

Further hospital characteristics evaluated were: staffing levels, exposure to competition, and treatment intensity (cost and treatments).

At a sub-provider level, some studies focused on clinician experience in terms of volume of stroke patients they see and clinician specialty. Other studies used a more investigatory approach, with multiple hypotheses about associations between performance against the indicators and a range of hospital factors, or even a range of patient and hospital factors.

4.6.3.3. Cross-sectional patient characteristic level

Similarly, some investigatory studies looked at how different patient groups performed against the indicators and, for some research, looked across a range of patient factors. Other studies investigated associations between performance and specific patient characteristics, such as
PART II: Stroke Care

ethnicity, age, gender, socio-economic status, distance to hospital, and insurance status.

Away from patient characteristics, other studies looked at the association between exposure of patients to different treatments, with a number of studies looking at the influence of day of admission on outcomes.

4.6.3.4. Temporal design

Under a third of studies (30/99) used longitudinal analysis, with a small selection using quasi-experimental design:

- regions, e.g. the effect of reimbursement changes for uncompensated care;
- hospitals, e.g. the association between quality and whether a hospital had been merged/acquired; and
- a single hospital, e.g. how introducing a more specialist service affected performance.

4.6.4. Details from datasets

4.6.4.1. Data source

The studies reviewed used a range of different administrative databases, with the most frequent including the:

- AHRQ-sponsored HCUP NIS and HCUP SIDs (see paragraph 2.1.1.2);
- National Patient Care Database's Patient Treatment File, which includes records of all hospital admissions from Veteran Association patients; and
- Medicare Provider Analysis and Review (MEDPAR) file, which includes data from claims for services provided to USA beneficiaries admitted to Medicare facilities, available since 1999.

Some studies used multiple administrative databases – for example, international studies for which there are no cross-national administrative databases – or, more commonly, administrative hospital data were supplemented with another source, such as mortality records to account for deaths outside the hospital setting. The administrative datasets were found to be heterogeneous with, for instance, many variations in the coding framework for operations/treatments received.
4.6.4.2. Case ascertainment

Almost three-quarters of studies used the ICD-9 framework for coding diagnoses, with only 11 being based on ICD-10. A third of studies did not describe which fields were used to identify strokes, although most that did used only the primary/principle diagnosis (57/66).

Limiting the description to those studies that used the ICD-10 framework, the most commonly used codes were: I61 (intracerebral haemorrhage); I63 (cerebral infarction); I64 (stroke, no specified as haemorrhage or infarction) and, to a lesser extent, I60 (subarachnoid haemorrhage) and I62 (other non-traumatic intracranial haemorrhage). However, there were numerous variations (Table 18). For instance, one study – described as evaluating all strokes – only used I63.237 In some instances, the studies set out prognostic differences between the stroke types as a rationale for making exclusions217,260 while in others the study was explicitly focused on a specific stroke type although no reason for this restriction was given.206,235

Table 18. Variations in use of ICD-10 diagnosis codes for identifying stroke

<table>
<thead>
<tr>
<th>ICD-10 code</th>
<th>Description</th>
<th>Examples of study</th>
</tr>
</thead>
<tbody>
<tr>
<td>H34.1</td>
<td>Central retinal artery occlusion</td>
<td>Zhu (2009)213; Field (2004)240</td>
</tr>
<tr>
<td>G45</td>
<td>TIAs and related syndromes</td>
<td>Zhu (2009)231</td>
</tr>
<tr>
<td>G45, not G45.4</td>
<td>TIA excluding transient global amnesia</td>
<td>Field (2004)219</td>
</tr>
<tr>
<td>I69</td>
<td>Sequelae of cerebrovascular disease</td>
<td>Katzenellenbogen (2010)293</td>
</tr>
<tr>
<td>all other I6x codes</td>
<td>See note*</td>
<td>Reklaitiene (2008)251; Slobbe (2008)224; Tu (2009)253</td>
</tr>
</tbody>
</table>

Examples of codes excluded

| I61         | Intracerebral haemorrhage | Saposnik (2007)202; Saposnik (2008)204 |
| I63.6       | Cerebral infarction due to cerebral venous thrombosis, non-phyogenic | Pajunen (2005)294 |
| I64         | Stroke, no specified as haemorrhage or infarction | Cho and Yun (2009)236 |

Notes: I60 – I64 and I69 described elsewhere, I65-6 (Occlusion and stenosis of precerebral and cerebral arteries, not resulting in cerebral infarction), I67 (Other cerebrovascular diseases), I68 (Cerebrovascular disorders in diseases classified elsewhere).
4.6.4.3. Extract validation

Few studies explicitly validated the accuracy of the underlying data or appropriateness of the assumptions used to extract details of stroke patients. Within the studies reviewed, the only coding items for which the validity was evaluated were the coding of stroke diagnoses and recording of thrombolysis including using comparison with non-administrative datasets. One study excluded hospitals for which the data was deemed of inappropriate accuracy.

4.6.5. Risk models

4.6.5.1. Case mix adjustment

The studies used various adjustments to mitigate the effects of differing case-mix. At its most rudimentary, some studies presented unadjusted rates when making comparisons and others stratified their findings by age, or age and gender. Some studies used modelling techniques to adjust for these latter two patient demographics. The most common set of variables used in the risk adjustment were age, gender and a measure of comorbidities however there were many additional variables considered in the studies (Table 19).

Table 19. Factors included in risk models

<table>
<thead>
<tr>
<th>Factor</th>
<th>Covariate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Patient characteristics</td>
<td>Social deprivation/socio-economic status 253,287,288</td>
</tr>
<tr>
<td></td>
<td>Ethnicity 300,232,283,284</td>
</tr>
<tr>
<td></td>
<td>Resident of country 237</td>
</tr>
<tr>
<td></td>
<td>Marital status 163,237,252</td>
</tr>
<tr>
<td></td>
<td>Distance to hospital 227</td>
</tr>
<tr>
<td></td>
<td>Insurance (payer, including Medicaid status) 229,236,237,239,244,247,262-264,272,281,284,289-291</td>
</tr>
<tr>
<td>Medical history</td>
<td>Place of admission 258 163,227,288</td>
</tr>
<tr>
<td></td>
<td>First vs recurrent stroke 301</td>
</tr>
<tr>
<td></td>
<td>Previous admissions: overall 206,220,271</td>
</tr>
<tr>
<td></td>
<td>CVD-specific 217</td>
</tr>
<tr>
<td></td>
<td>Stroke-specific 263,301</td>
</tr>
<tr>
<td></td>
<td>TIA-specific 163,301</td>
</tr>
<tr>
<td>Stroke/treatment details</td>
<td>Stroke type 163,213,239,252,263,279,286,302,215,246</td>
</tr>
<tr>
<td></td>
<td>Temporal: year of treatment 247 216,264,274,283,290</td>
</tr>
<tr>
<td></td>
<td>season 290,271,291,293,298,302,277</td>
</tr>
<tr>
<td></td>
<td>proportion of patients admitted at weekends 271</td>
</tr>
<tr>
<td></td>
<td>Signs of cognitive functions and awareness 237,228,264,272,273,281,213,264,274,279,223,264,274,279</td>
</tr>
<tr>
<td></td>
<td>certain surgical operations 213,264,274,279,286,287,288,290</td>
</tr>
<tr>
<td></td>
<td>hemiplegia or hemiparesis 213,264,274,279,286,287,288,290</td>
</tr>
<tr>
<td></td>
<td>residual neurologic deficits 213,264,274,279,286,287,288,290</td>
</tr>
<tr>
<td></td>
<td>Craniotomy 277,274,279,286,287,288,290</td>
</tr>
<tr>
<td></td>
<td>Other conditions (e.g. pneumonia, atrial fibrillation, urinary incontinence) 237,228,264,272,273,281,213,264,274,279,223,264,274,279</td>
</tr>
<tr>
<td></td>
<td>Severity (late effects of stroke, which includes cognitive deficits. Speech and language deficits an, and paralysis, dysphagia, incontinence, and delirium) 271,228,264,272,273,281,213,264,274,279,223,264,274,279</td>
</tr>
<tr>
<td></td>
<td>Gastronomy, tracheotomy, interventional intracranial thrombolysis, interventional angioplasty, endarterectomy, operative angioplasty, extracranial-</td>
</tr>
</tbody>
</table>
Chapter 4: Stroke literature review

4.6.5.2. Participant exclusions

To account for potential bias from different stroke, patient and hospital characteristics, studies used a range of exclusion criteria to define their participants. The most common criterion was age of patient, although sometimes this was necessitated by the underlying administrative databases being age-specific. Cut-offs for the exclusion of young people included 18, 276 20, 286, 299 35, 224 and 40. 238 Conversely, other publications excluded older patient, such as those over 105 years old, 286 with one study excluding those over 75 because of the possibility of increasing unreliability of the clinical diagnoses of stroke with advancing age. 303

Other exclusions related to removing patients with atypical prognosis, such as: lengths of stay over one year; 220 previous admission admitted for CVD in five years before admission; 220, 286 patients admitted from long-term care; 257 208 non-residents; 208 transfers; 208, 287 admissions not through emergency department; 269, 286 patients discharged within 1 day of admission or left the hospital against medical advice. 215, 265 Other studies excluded hospitals with low stroke numbers 227, 298 or patients for whom diagnosis and treatment codes suggested an alternative primary diagnosis to stroke. 276

4.6.5.3. Summary of key factors

The following exhibit (Table 20) summarises how different studies have controlled for, or examined, some potential determinants of variation in quality and safety. This incorporates details about the key dependent variables, case-mix adjustment and exclusions from the indicator denominators.
### Table 20. Summary of key factors used in studies

<table>
<thead>
<tr>
<th>Level</th>
<th>Key covariate</th>
<th>Risk-adjusted</th>
<th>Participant exclusions</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Regional</strong></td>
<td>Regional networks</td>
<td>Camilo (2005)250</td>
<td>e.g. transferred case in Mitchell (1996)227</td>
</tr>
<tr>
<td></td>
<td>Competition</td>
<td>e.g. Volpp (2005)288, 39</td>
<td>Tung (2009)268</td>
</tr>
<tr>
<td><strong>Hospital</strong></td>
<td>Location</td>
<td></td>
<td>e.g. Cross (2003)239, 262</td>
</tr>
<tr>
<td></td>
<td>Size (volume)</td>
<td>e.g. Bardach (2002)25, 54-58</td>
<td>e.g. Gillum (2008)269, 291</td>
</tr>
<tr>
<td></td>
<td>Size (beds)</td>
<td>e.g. Andaluz (2008)24, 52, 53</td>
<td>e.g. Polanczyk (2002)47, 53, 71, 94</td>
</tr>
<tr>
<td></td>
<td>Teaching status</td>
<td>e.g. Andaluz (2008)21, 21, 235, 2</td>
<td>e.g. Saposnik (2007)12, 260, 263, 268, 275, 276</td>
</tr>
<tr>
<td></td>
<td>Comparison/clustering</td>
<td>e.g. NHS Information Centre (2011)42, 43</td>
<td>e.g. Volpp (2005)38, 39, 51</td>
</tr>
<tr>
<td></td>
<td>Ownership</td>
<td></td>
<td>e.g. Polanczyk (2002)11, 268</td>
</tr>
<tr>
<td></td>
<td>Specialist centre</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Clinician</strong></td>
<td>Specialty</td>
<td>e.g. Mitchell (1996)237, 229, 269, 271</td>
<td>e.g. Saposnik (2008)56, 58, 78</td>
</tr>
<tr>
<td><strong>Stroke</strong></td>
<td>Type</td>
<td>e.g. Borzecki (2010)209, 218, 258, 2</td>
<td>e.g. Carinci (2007)42, 58, 73, 83, 90, 92, 115, 117</td>
</tr>
<tr>
<td></td>
<td>Year</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Day of stroke (weekend)</td>
<td>e.g. Bell (2001)253, 268, 289</td>
<td>Howrey (2011)271</td>
</tr>
<tr>
<td><strong>Stroke outcome</strong></td>
<td>Length of stay</td>
<td></td>
<td>e.g. Kind (2007)272</td>
</tr>
<tr>
<td></td>
<td>Discharge destination</td>
<td></td>
<td>238, 273, 277, 281</td>
</tr>
<tr>
<td></td>
<td>ICU</td>
<td></td>
<td>Reed (2001)275</td>
</tr>
<tr>
<td><strong>Patient</strong></td>
<td>Socio-economic status</td>
<td>e.g. Gregory (2009)239, 286</td>
<td>e.g. Lee (2007)277, 261, 275</td>
</tr>
<tr>
<td><strong>characteristics</strong></td>
<td>Ethnicity</td>
<td>e.g. Lee (1998)278, 279</td>
<td>e.g. Bell (2001)45, 93, 94</td>
</tr>
<tr>
<td></td>
<td></td>
<td>220, 232, 280-284</td>
<td></td>
</tr>
</tbody>
</table>
Chapter 4: Stroke literature review

4.6.7. Details from indicators

4.6.7.1. Range of indicators

Across the 99 papers, 166 (not necessarily distinct) indicators were described – an average of 1.7 indicators per paper – with no study having more than 5 indicators. Across the publications there were 20 categories of indicators, which cover: urgent acute care; acute treatment; complications; mortality; discharge; review; post-discharge (Table 21, with a full list given in Appendix D, p304). Through conducting the review, I also identified potential indicators which had yet to be applied to administrative data, such as ‘readmission for incontinence’, 160 and ‘newly institutionalised within 2 weeks of admission’. 304 Mortality, readmission and thrombolysis rates were all often described across the published literature. In line with the inclusion criteria, ‘hospital charges’ was not regarding as a quality indicator, even though this was common within the included studies 211,227,237,247,273,275,288,290. Some studies reported compound measures such as: ‘complications’ which covered either pneumonia, UTI and DVT 243 or pneumonia, PE, DVT 219; or ‘adverse outcome’ which combined in-hospital death with discharge destination other than home 284.

4.6.7.2. Indicator validation

Approaches used to take assurance on the validity of the indicator results were both sparse and varied. One of the studies compared performance across different indicators constructed from the same dataset; however, in this instance, few pairs had significant correlation. 222 A more common approach was to compare the results from administrative data to another data set, with the majority looking at mortality rates 240,248,283,285,303 with the other data sets ranging from death registers, to other administrative database 274 and case note reviews at one of the hospitals. 228 One study used a present-on-admission (POA) flag within the administrative data to see the effect on the sensitivity and specificity of the resulting indicator. 298

The most common form of validation described in the reviewed publications was on the risk-adjustment, with 16 studies using different sets of variables to create the various models for case-mix adjustment within their studies 163,166,209,212,216,227,233,239,242,249,257,272,273,282-284 and another study using competing analytical techniques for modelling risk. 269 Apart from these studies evaluating risk-adjustment, few of the remaining studies explicitly reported the model fit. 226,233,242,244,268,289 One
study investigated the impact of missing data and patient identifiers and two repeating the analysis on a different stroke extract (derived from different sets of stroke codes).

**Table 21. Categories of measures**

<table>
<thead>
<tr>
<th>Pathway</th>
<th>Outcome measures</th>
<th>Indicator type</th>
<th>Example paper</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Urgent acute care</strong></td>
<td>Thrombolysis</td>
<td>Process</td>
<td>Adeoye (2011)</td>
</tr>
<tr>
<td></td>
<td>Urgent brain scan within 24 hours</td>
<td>Process</td>
<td>Lazzarino (2011)</td>
</tr>
<tr>
<td></td>
<td>Emergency hospital transfers</td>
<td>Structure</td>
<td>Kind (2007)</td>
</tr>
<tr>
<td><strong>Acute treatment</strong></td>
<td>Depression and other mental health problems following stroke</td>
<td>Outcome</td>
<td>Williams (2004)</td>
</tr>
<tr>
<td></td>
<td>Inpatient rehabilitation services</td>
<td>Process</td>
<td>Kapral (2002)</td>
</tr>
<tr>
<td></td>
<td>Warfarin use</td>
<td>Process</td>
<td>Smith (2005)</td>
</tr>
<tr>
<td></td>
<td>Length of stay</td>
<td>Outcome</td>
<td>Tu (2003)</td>
</tr>
<tr>
<td><strong>Complications</strong></td>
<td>Urinary Tract Infections (UTIs)</td>
<td>Outcome</td>
<td>Censullo (2008)</td>
</tr>
<tr>
<td></td>
<td>Pneumonia due to swallowing problems</td>
<td>Outcome</td>
<td>Field (2004)</td>
</tr>
<tr>
<td></td>
<td>Lower respiratory tract infection (LRTI)</td>
<td>Outcome</td>
<td>Laudicella (2008)</td>
</tr>
<tr>
<td></td>
<td>Deep Vein Thrombosis</td>
<td>Outcome</td>
<td>Censullo (2008)</td>
</tr>
<tr>
<td></td>
<td>Pulmonary Embolism</td>
<td>Outcome</td>
<td>Field (2004)</td>
</tr>
<tr>
<td></td>
<td>Clostridium difficile</td>
<td>Outcome</td>
<td>Laudicella (2008)</td>
</tr>
<tr>
<td></td>
<td>Adverse outcome</td>
<td>Compound</td>
<td>Zacharia (2010)</td>
</tr>
<tr>
<td><strong>Mortality</strong></td>
<td>Craniotomy: mortality rate</td>
<td>Outcome</td>
<td>Borzecki (2010)</td>
</tr>
<tr>
<td></td>
<td>In-patient death within 30 days of admission</td>
<td>Outcome</td>
<td>Ostbye (1997)</td>
</tr>
<tr>
<td></td>
<td>All-location deaths within 30 days of admission</td>
<td>Outcome</td>
<td>Tirschwell (1999)</td>
</tr>
<tr>
<td><strong>Discharge</strong></td>
<td>Return to normal place of residence</td>
<td>Outcome</td>
<td>Votruba (2006)</td>
</tr>
<tr>
<td><strong>Post-discharge</strong></td>
<td>Emergency readmission to hospital following treatment from a stroke (all cause)</td>
<td>Outcome</td>
<td>Kamalesh (2007)</td>
</tr>
<tr>
<td></td>
<td>Recurrent stroke within 28 days</td>
<td>Outcome</td>
<td>Tu (2003)</td>
</tr>
</tbody>
</table>
4.7. Discussion

4.7.1. Principle findings
The literature review identified 99 studies that discussed the use of administrative data to measure the quality or safety of hospital stroke care and found 166 indicators, with the most prevalent relating to mortality, readmissions and thrombolysis rates. There has been a general increase in the application of these indicators, with over two-thirds of the studies using data since 2000.

4.7.2. Source of data
Three-quarters of the studies were from North America (USA or Canada) who use ICD-9 and a combination of ICD-9 and ICD-10 diagnosis codes, respectively. As such, only a small minority of studies, equating to 12%, use ICD-10 diagnosis codes. There is no one-to-one mapping between the two frameworks although many of the codes have equivalents. This, however, highlights that application of the measures and findings cannot be generalised everywhere. The reason for the prominence of research within North America may be an artefact of a greater use of established administrative datasets in these countries and, also, a lack of national audit data in USA (to act as an alternative data source).

4.7.3. Stroke case ascertainment strategies
4.7.3.1. Diagnosis fields
The review identified studies that used all diagnoses fields, and one that used first to fifth diagnosis positions, with these strategies likely to increase the sensitivity but reduce specificity of the algorithm for identifying strokes within the administrative database. However, among those studies that did describe which diagnosis fields were interrogated, the majority used only the primary/principle diagnosis, which is in keeping with previous findings on the use of administrative databases for quality research. Within Australian administrative health data, the primary diagnoses are rarely missing and stroke diagnoses appearing in administrative data sets are accurate 93% of the time. Similar results have been found with ICD-9 databases. Moreover, other studies have added that using the first diagnosis field only increases the probability that the patient had suffered a stroke on the date of admission. This is also consistent with coding guidance in the UK which advises that acute stroke should be coded in the primary diagnosis field.

4.7.3.2. Stroke diagnosis codes
From reviewing the literature, it was also evident that there was no consensus over what diagnosis codes were used in the studies on administrative data. As an example, looking just at the research
based on ICD-10 databases, the comprehensiveness of the strategy for identifying stroke admissions range from using just one 3-digit group of diagnosis code (I61) to using ten (I60 – I69). This is significant since certain stroke types have distinct care pathways. For instance, ischaemic strokes can potentially be treated using thrombolytic drug treatment; however, this would result in an adverse outcome if administered in treating a haemorrhagic stroke. Similarly, subarachnoid haemorrhagic strokes will likely receive substantially different urgent care to the other emergency stroke types.

A 2005 study found that stroke coding was equally good for both ICD-9 and ICD-10 and while there were some differences in coding of the different types of stroke, these variations were not statistically significant and the most atypical coding performance was not for actual stroke (rather for TIs). However, not all strokes will be either accurately diagnosed or, even where they are, coded correctly or specifically enough. As a result, there is variable use of the unspecified or ill-defined stroke codes (e.g. I64 in ICD-10 and 436 Acute, but ill-defined, cerebrovascular disease in ICD-9). Therefore, a decision has to be made whether to use more limited codes to identify strokes, e.g. excluding those recorded as subarachnoid haemorrhages, which will increase the specificity of the denominator but might increase bias resulting from variable use of the unspecified stroke codes.

One study specifically looked at the effect of using two different case ascertainment strategies on an outcome indicator (mortality). The risk-adjusted models for cohorts produced from the highly specific and, separate, highly sensitive sampling strategies were both robust and had some impact on respective rankings of the regional networks against the outcome measure, and whereas two outliers were identified in one cohort, none were identified in the other.

4.7.4. Indicator validity

4.7.4.1. Risk-adjustment

The importance of adjusting for case-mix in research on stroke care using administrative data is longstanding. However, continuing the theme of the findings of this review, there was unexplained variations in the approaches to adjusting for stroke type and both patient- and hospital-characteristics. Studies have found that adjusting for socio-economic status is inadequate without more comprehensive measures being included; however, even this level of risk-adjustment is more sophisticated than that applied in many of the studies. Likewise many studies did not adjust for comorbidities despite research showing that it is useful for predicting risk-adjusted in-hospital case-fatality in stroke outcome studies.
4.7.4.2. Coding of covariates, activity and outcomes

Accurate coding – including of patient and hospital characteristics, treatment activity and outcomes – is inherently important if the data are to be used to identify true variations in the quality and safety of care. A number of studies evaluating the accuracy of coding for stroke patients – looking at variables such as coding of comorbidities\textsuperscript{310,312,313} and nonmedical and socio-demographic data elements\textsuperscript{269,314} – were identified through the initial search of citations and subsequent review of bibliographies and, in general they showed that the coding was of reasonable accuracy.\textsuperscript{310} However, the results are neither appropriate for synthesising – since heterogeneity between the studies mean that it is not possible to undertake any meta-analysis of the results – nor generalisable – since there are likely differences between the coding practice in different countries and changes since these non-contemporaneous studies were published.

One commonly utilised indicator that directly relies on coding completeness is thrombolysis rate. A number of studies on this topic that were referenced in the reviewed studies with, in general, high specificity and mixed results on sensitivity, although again these are not necessarily generalisable due to the heterogeneity of databases, differing national incentives to complete them correctly and limited coverage of some of these studies.\textsuperscript{225,295,296}

4.7.4.3. Indicator validation

As well as the ‘upstream’ validation of the case ascertainment strategies and robustness of the risk-adjustment model, the literature review revealed some examples of ‘downstream’ explicit validation of the indicator results, for instance by comparing the results with indicators derived from other sources. Despite the existence for various frameworks of taking assurance on the validity of such indicators,\textsuperscript{130,315} none of the studies described a comprehensive validation strategy, certainly not in comparison to this study’s analytical framework (Table 8, p67).

4.7.5. Strengths and limitations

4.7.5.1. Strengths

This study represents the most comprehensive review of measures of the quality and safety of stroke care using administrative data. An obvious strength of this literature review process which focused on a particular specialty is that it seems to be a sensitive strategy for identifying relevant research and, as such, provides more assurance that the resultant citations are comprehensive. For instance, the literature review followed a similar process to a review of all patient safety measure based on routinely-collected hospital data which identified only a similar magnitude of studies (n = 124) despite its far broader scope.\textsuperscript{67}
4.7.5.2. Limitations of the review process

The limitations of this review process include the potential bias introduced from the use of a single reviewer and exclusion on non-English language articles. The exclusion of non-English articles was consistent with a previous review of patient safety measures, which found that such citations were limited in number. No formal meta-analysis could be undertaken because of the heterogeneity in study methodologies and underlying data frameworks. I did not use a template to assess the quality of the papers since there was a very diverse range of purposes across the studies and, as already discussed, validity needs to be proportionate to the purpose. Given the strong evidence on the existence of publication bias, studies using indicators which did not identify significant results when assessing the quality and safety of stroke care may have been missed by the search strategy in this study.

4.7.5.3. Limitation of review protocol

The literature review again reiterated the differences in extant administrative data and how they are used, such as in data-linkage studies. This makes the formulation and application of strict inclusion/exclusion criteria for such literature review difficult. In keeping with the review criteria for this research as described earlier, a study based on Uniform Data System for Medical Rehabilitation (UDSMR) was excluded, since the indicators in the study used this specific rehabilitation data collection only and were not readily reproducible with other administrative data sources, whereas a study which linked this to the Medicare Provider and Analysis and Review files (MedPAR) administrative database was included, since the measure described could be recreated. Similarly, a study with measures using linked data with variables including physical activity (not commonly included in administrative datasets) was excluded; however, a study which supplemented the hospital administrative data with clinical register data was included since the measures could still be replicated - albeit with some minor amendments - using purely administrative data.

4.7.6. Conclusion

4.7.6.1. Contribution to knowledge of indicators

Looking at a specialty in detail, had advantages over a broader scope; not least, to better identify existing indicators. The study has highlighted a paucity of use of such indicators and a lack of transparency and consensus on key assumptions.
4.7.6.2. **Recommended uses**

The studies in the review have been used for a range of purposes, from comparing performance across countries to evaluating the impact of the implementation of a specialist stroke unit within a single hospital. Administrative data do not include all the process measures collected in bespoke registers, such as provision of anti-thrombolytics, however they cover the main outcomes – such as including all three key outcomes (length of stay, mortality and discharge destination) from one register.\(^{320}\) And indeed, in some cases, registers had to be linked to administrative data to get details of areas such as clinician experience.\(^{255}\) However, continued monitoring of stroke mortality has been described as “essential”\(^ {321,322}\) and, more generally, surveillance of stroke care is necessary to understand and meet the demand for stroke care\(^ {323}\) with the current lack of measures having been cited as partly attributable to the limitations on the adoption of best practice.\(^ {324}\)

As a result, administrative data has been recommended as a tool for identifying potential quality problem deserving further review.\(^ {310,323,325}\) Moreover, there is some evidence to suggest that such measures can bring meaningful change with, for instance, a decline in stroke mortality rates associated with the introduction of an initiative to measure hospital performance, although this was not statistically significant (p=0.17).\(^ {326}\) While the literature suggests there is a role for stroke-specific indicators based on administrative data, there is scope for the current raft of measures to be improved in their construction and application. Table 22 sets out six key recommendations for future stroke research based on the findings of the review, which are enacted in the following three chapters.
### Table 22. Summary of implications for study design

<table>
<thead>
<tr>
<th>Finding</th>
<th>Implication</th>
<th>Recommendation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Despite the recommendation of using indicators in passive surveillance of stroke services, the majority of research has been restricted to North America, and using ICD-9 databases.</td>
<td>Barriers to generalising findings across borders mean that little is known about the use of such indicators in other countries, especially those using ICD-10.</td>
<td>Key study designs should be replicated using data from outside USA or Canada to identify any potential outliers.</td>
</tr>
<tr>
<td>Two-thirds of studies applied only one indicator.</td>
<td>Overall level of quality of care at a provider-level cannot be deduced from the results of only a small set of indicators.</td>
<td>To use the depth of information within the administrative data to construct more comprehensive set of indicators.</td>
</tr>
<tr>
<td>There is a lack of consensus and transparency about the assumptions used in applying the indicators to administrative data</td>
<td>Barrier to generalising, validating or benchmarking results.</td>
<td>To explicitly set out the assumptions used in studies, such as the choice of diagnosis codes and fields.</td>
</tr>
<tr>
<td>While concerns remain over bias due to differences in coding practice, none of the literature reviewed explicitly applied sensitivity analyses on effect of variations in coding.</td>
<td>The effect of variation in coding practice is unknown.</td>
<td>To analyse the potential effect of coding variation.</td>
</tr>
<tr>
<td>To consider areas known to be potentially important and where comparisons are less affected by coding practice (e.g. weekend care)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Less than a third of studies looked at trends in performance over time.</td>
<td>Potential shortcomings in quality and safety are not being identified.</td>
<td>Research into the potential opportunities for monitoring quality and safety using longitudinal designs.</td>
</tr>
</tbody>
</table>
Chapter 5.
Stroke variation and validation
PART II: Stroke Care

Overview

Context
The review of stroke literature described in the previous chapter revealed a paucity of studies evaluating stroke services using administrative data outside Northern America, despite suggestions that this would be a useful surveillance tool. Across the research, there was a lack of transparency of the assumptions used with little use of the potential richness of indicators available or opportunities to validate results. In this chapter I aim to address these limitations by examining the feasibility of evaluating stroke care in England using routinely collected administrative data.

Methods
Six indicators, spanning the hospital care pathway, were applied to English data in 2009/10. Logistic regression analyses were used to adjust for case-mix, and funnel plots to look at potential outliers. I investigated the effect of variations in hospitals’ data coding practice.

Findings
The indicators were sufficiently sensitive to identify significant variations in performance as recorded in the data, with 181 occurrences of hospitals performing statistically differently than expected at the 99.8% significance level across the six indicators. Differences in coding practice appeared to only partially explain the variation. The work showed that administrative data can be used to indicate potential concerns around quality and safety, in specific areas, across the whole stroke care pathway. The chapter discusses work – on demographic differences – based on one of the indicators which emphasises the impact of stroke care and suggests that there is scope to improve.

What this chapter adds

| Most comprehensive assessment of stroke services using administrative data in any country |
| The most sophisticated evaluation of validity of any stroke assessment using administrative data |

Acknowledgements and related papers

Advice was provided by: Sue Eve-Jones (UK Professional Association of Clinical Coders) on clinical coding; and Charlie Davie provided advice on clinical implications. Antonio Lazzarino undertook the data extraction and analysis for the ethnographic study (reported in the discussion).


Chapter 5: Stroke variation and validation

5.1. Background

5.1.1. Stroke care in England

5.1.1.1. Priority of stroke care measurement

Stroke costs the National Health Service (NHS) in England £3 billion a year in direct care costs, and is the largest cause of adult disability. In July 2010 the Government set out its policy intentions for the NHS including to measure performance “against results that really matter”, exemplifying this with stroke survival rates and emergency readmissions to hospital within 28 days of discharge for stroke. The proposals also reiterated that, compared with other countries, the NHS has achieved relatively poor outcomes in some measures of stroke and, as such, stroke was earmarked as one of the two areas with the most scope for improvement. More recently, the NHS’s operating framework for 2011/12 highlighted the scope for improving stroke outcomes, in particular through access to timely scanning and, if clinically indicated, thrombolysis. In recent years, the Department of Health in England have also introduced payments for patients receiving thrombolysis, early scans, and being admitted directly onto stroke units.

A national report also recently highlighted both the enduring variations in hospital care and that some levels of care remain unacceptably low. This phenomenon is not restricted to England, and as a result research has also called for indicators to measure stroke care performance.

5.1.1.2. Existing measurement of stroke in England

To date, the most comprehensive and high-profile study to benchmark stroke care in English hospitals has been the National Stroke Sentinel Audit (NSSA). The NSSA, which was first run in 1998, benchmarks hospitals’ stroke care over both organisational and clinical domains, based on a survey of hospital stroke teams and medical record review of around 60 consecutive patients. A further bespoke data collection was required after the introduction of mandatory measures regarding stroke care in the NHS operating framework for 2008-09 to 2010-11. Since 2000, the National Centre for Health Outcomes Development (NCHOD) has published a compendium of some 300 indicators, including three indicators based on administrative data specific to hospital stroke care (number of admissions, readmissions and mortality), with a further six measures relating to stroke prevention.
5.1.2. Revisiting limitations in existing literature

The review of literature of stroke measures suggested that, to address the shortcomings of existing research, future efforts should focus on (excluding the recommendations relating to temporal study designs which are covered in the following chapter):

1. applying key study designs for surveillance outside the USA or Canada to identify potential quality issues;
2. using a more comprehensive set of indicators than has been done to date;
3. being explicit about the assumptions used; and
4. understanding the potential bias from variation in coding practice.

5.2. Objectives

In this chapter I investigate the feasibility of using HES data to evaluate the quality and safety of stroke care at a hospital level focusing on the ability of measures to identify important effects in a robust fashion.

5.3. Methods

5.3.1. Selecting indicators

5.3.1.1. Selection criteria revisited

Recall the criteria listed earlier (paragraph 3.5.1.2, p76) for the selection of indicators. For this set of studies on the stroke specialty, the experts involved in getting a consensus on the indicators are listed in paragraph 5.3.2.1, p119.

5.3.1.2. Selection of indicators

The literature review described in the previous chapter identified 20 broad categories of indicators with the potential to be applied to administrative data in England. Following the selection criteria, a subset of six indicators was chosen for application here. The indicators were chosen to cover the hospital stroke medical care pathway and include: process measures (e.g. scanning rates); outcome measures that are proxies for certain aspects of care (e.g. pneumonia rates, with higher rates indicating fewer patients receive swallow assessments); and outcome measures (e.g. emergency readmission rates) (Figure 9).
5.3.2. Developing indicator definitions

5.3.2.1. Development process

The selected indicators were then refined using a multi-faceted approach.

1. I derived draft definitions using details from algorithms applied in existing literature, identified through the systematic review outlined in the previous chapter.

2. I used an iterative consensual approach to ascertain feedback from a multidisciplinary panel of expert, consisting of a senior stroke physician (who was also involved in the indicator selection), a public health doctor, a clinical coding specialist (who had previously worked on developing indicator definitions) and administrative data experts (from Dr Foster Intelligence and Imperial’s Dr Foster Unit).

3. The revised definitions were published online – through the Health Service Journal website – as part of an open consultation on clinical indicators. 332

4. Where necessary the panel of experts were consulted on addressing comments raised on the definitions from the open consultation.

5.3.2.2. Final definitions

The finalised indicator definitions are included below in Figure 10.
**Figure 10: Details of stroke indicators used in analysis**

<table>
<thead>
<tr>
<th>Indicator</th>
<th>Rationale</th>
<th>Numerator (discharges meeting the inclusion/exclusion criteria for the denominator with...)</th>
<th>Exclusions</th>
</tr>
</thead>
<tbody>
<tr>
<td>a Same day and by-next-day scanning</td>
<td>Brain scanning should be performed immediately (when indicated) or as soon as possible (NICE 2008). Hospitals receive incentive payment for immediate scanning.</td>
<td>OCPS codes for CT or MRI brain scan (U05.1/2 or U21.1/2 with Z01.9)</td>
<td>Patients who die on day of admission (same day scan) or within day (by-next-day scan)</td>
</tr>
<tr>
<td>b Thrombolysis</td>
<td>Measure of provision of thrombolysis. Hospitals receive payment for providing thrombolysis.</td>
<td>OPCS codes for thrombolysis (fibrinolytic drugs, X83.3)</td>
<td>Patients outside thrombolysis licence age range of 18 – 80</td>
</tr>
<tr>
<td>c Aspiration pneumonia</td>
<td>Aspiration pneumonia is an indicator of no swallow assessment or related care</td>
<td>ICD-10 codes for aspiration pneumonia: J69.0 (due to food and vomit) and J69.8 (other solids and liquids)</td>
<td>Pneumonia codes recorded in episodes that end before the first episode with a stroke diagnosis</td>
</tr>
<tr>
<td>d 30-day in-hospital mortality</td>
<td>Some mortality following stroke is potentially avoidable</td>
<td>Flag for death at discharge and length of stay after stroke &lt;30</td>
<td>None</td>
</tr>
<tr>
<td>e Discharge to usual place of residence within 56 days</td>
<td>Proxy measure for successful outcome of rehabilitation and availability of on-going care</td>
<td>Length of stay after stroke &lt;56 days, and admission source and discharge destination suggesting return to usual place of residence</td>
<td>Admissions which end in death</td>
</tr>
<tr>
<td>f 30-day emergency readmissions (all cause)</td>
<td>Some readmissions are potentially avoidable</td>
<td>Emergency admissions within 0-29 days of discharge</td>
<td>Admissions which end in death</td>
</tr>
</tbody>
</table>

### 5.4. Application and validation of indicators

#### 5.4.1. Assumptions

The next stage was to extract records of stroke admissions from 1 April 2009 to 31 March 2010. The assumptions used in the algorithm for identifying strokes in this study - based on previous studies, consultation with clinical coders and review of coding guidance - are listed in Figure 11. The assumptions include: diagnosis codes, fields and type; identification of the date of the stroke; treatment of transfers, including linkage into superspell admissions and assigning a responsibility to the hospital; and treatment of in-hospital strokes.
### Figure 11: Assumptions for identifying stroke using Hospital Episode Statistics

<table>
<thead>
<tr>
<th>Issue</th>
<th>Assumption</th>
<th>Rationale</th>
<th>Alternatives</th>
<th>Implication</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Diagnosis codes</strong></td>
<td>ICD-10 codes I60 – I64</td>
<td>Review of literature/clinical coding advice</td>
<td>ICD-10 codes G08.X, G45 – 46, H34, I65 – 69, Q25.8, Q28, R47 have also been used to identify patients in stroke studies</td>
<td>Incorrect stroke numbers: the chosen criteria are quite strict, in comparison with most studies</td>
</tr>
<tr>
<td><strong>Diagnosis field</strong></td>
<td>HES primary diagnosis field only</td>
<td>Review of literature/clinical coding advice and guidance</td>
<td>Some or all of the secondary diagnosis codes</td>
<td>Underestimate total number of strokes: those incorrectly coded in the secondary fields are excluded</td>
</tr>
<tr>
<td><strong>Defining stroke type</strong></td>
<td>Strokes categorised by 4-digits of first stroke diagnosis code meeting inclusion criteria (above)</td>
<td>Subsequent stroke codes could refer to secondary strokes</td>
<td>[1] Final stroke diagnosis; [2] Supersede coding of I64.X (not specified) if another stroke code is listed; [3] Three-digit codes</td>
<td>Incorrect classification of stroke: where stroke is subsequently re-diagnosed, strokes labelled under the initial, provisional diagnosis. This will affect case mix adjustments</td>
</tr>
<tr>
<td><strong>Date of stroke</strong></td>
<td>Start date of first spell with stroke diagnosis</td>
<td>Coding guidance</td>
<td>[1] Start date of episode with first stroke diagnosis; [2] Start date of admission</td>
<td>Incorrect stroke date: In-hospital strokes not recorded as a new spell may have early stroke date. Strokes not diagnosed in the first spell of admission may have late stroke date</td>
</tr>
<tr>
<td><strong>Linking spells</strong></td>
<td>Episodes of care linked to same superspell</td>
<td>Reduce double-counting</td>
<td>Analyse by spell rather than superspell</td>
<td>Incorrect stroke numbers: Superspells with &gt;1 spell are not counted more than once</td>
</tr>
<tr>
<td><strong>Responsible hospital</strong></td>
<td>First hospital of admission</td>
<td>First hospital are responsible for onward transfers</td>
<td>Hospital providing majority of (e.g. most bed days) or key aspects of care</td>
<td>Performance assigned to wrong hospital: if local pathways involve hospital transfers, the performance of the initial hospital will be influenced by the care of subsequent hospital</td>
</tr>
<tr>
<td><strong>Type of admission</strong></td>
<td>Emergency and elective admission</td>
<td>To not exclude in-hospital strokes</td>
<td>Emergency admissions only</td>
<td>Include non-acute stroke admissions as some episodes with primary diagnosis of stroke might relate to treatment for a previous stroke</td>
</tr>
</tbody>
</table>
5.4.2. Risk-adjustment

5.4.2.1. Overview of case-mix adjustment

Once an extract had been obtained using the above criteria, the first step was to identify outliers, which involved applying the indicator definitions to obtain denominators and numerators for each hospital. A logistic regression was used to calculate an expected number of numerator events based on the case mix for each hospital to account for: age, sex, socio-economic deprivation quintile, number of previous admissions, co-morbidities (Charlson index), month of discharge, ethnic group, source of admission (including whether admitted as an emergency or elective patient), and stroke type (4-digit ICD-10 diagnosis code). Charlson index was fitted as a continuous variable, with all others fitted as categorical variables. The risk model would not converge for some specifications and, for these, admission source was omitted from the case-mix adjustment. The methods used are described in more detail in paragraphs 3.6.2 to 3.6.6.

5.4.2.2. Reporting of process measures

For most process indicators, risk adjustment plays a smaller role than it does for outcome measures. Process measures (scanning and thrombolysis) are reported here as unadjusted (crude) rates, although adjusted rates were also calculated to check the validity of this approach.

5.4.3. Statistical testing and validation

Crude and standardised rates were plotted using funnel plots with 95% and 99.8% control limits and identified outliers (paragraph 3.6.8, p81). Hospitals’ performance across the different indicators was compared to evaluate our hypothesis that certain indicators would be correlated. This involved 21 pairwise comparisons (the scanning measure was considered as two separate measures – same day and by next day scans), with further detail on this method included in the discussion below. Since these were hypothesised as being correlated and not a un-hypothesised statistical trawling exercise, Bonferroni correction or other adjustment to p-values for multiple comparisons was not used; however, I do report which correlations are significant at the 99.8% level.

5.4.4. Sensitivity analysis for coding practice

There is a risk that hospitals’ performance for these indicators might be largely affected by variation in the way they code their data rather than due to differences in quality and safety. As such, the final stage of analysis was to investigate, at a hospital level, the consistency of coding practice and evaluate the relationship between any coding bias and hospital performance. The three coding practices investigated, by comparing to performance in a relevant indicator, were:
[1] ‘coding depth’, with the coding practice in some hospitals increasing the likelihood that they record secondary diagnoses and, therefore, identify complications and comorbidities;

[2] use of the ICD-10 diagnosis code I64 - Stroke, not specified, since it may bias the risk adjustment for the outcome measures; and

[3] date of stroke, since some hospitals may be more likely to assign a stroke code to the episode when the stroke was formally diagnosed whilst others would assign a stroke code to the date of admission.

For the latter practice, I calculated the average time-lags, by hospital, between the date used to signify the stroke in this study (start date of the first spell with a stroke diagnosis) and both the start date of the first episode with a stroke diagnosis and the start date for the superspell.

The effect on the sensitivity of the measures when including a hospital-level variable to account for variations in coding practice was investigated by fitting generalised linear models and again plotting on a funnel chart to identify outliers.
5.4.5. Alignment with analytical framework

Table 23. Analytical framework

<table>
<thead>
<tr>
<th>Domain</th>
<th>Study design</th>
<th>Method described</th>
<th>Results described</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Impact</td>
<td>Descriptive, national rates</td>
<td>-</td>
<td>5.5.1.1</td>
</tr>
<tr>
<td>2. Amenable</td>
<td>Descriptive, range</td>
<td>-</td>
<td>5.5.1.2</td>
</tr>
<tr>
<td></td>
<td>Demographic-level variation</td>
<td></td>
<td>5.6.2</td>
</tr>
<tr>
<td>Importance</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. Face and content validity</td>
<td>Consultation</td>
<td>5.3.2.1</td>
<td>5.3.1</td>
</tr>
<tr>
<td>4. Precision</td>
<td>Hospital results displayed on funnel plots</td>
<td>5.4.3</td>
<td>5.5.2</td>
</tr>
<tr>
<td>Scientific soundness of indicators</td>
<td>Correlations between the set of indicators, including outcome v. process indicators</td>
<td>5.4.3</td>
<td>5.5.3</td>
</tr>
<tr>
<td>Scientific soundness of indicators</td>
<td>Risk-adjustment model</td>
<td>5.4.2</td>
<td>5.5.5</td>
</tr>
<tr>
<td></td>
<td>Sensitivity analyses of influence of coding error and practice.</td>
<td>5.4.4</td>
<td>5.5.4</td>
</tr>
<tr>
<td>7. Data availability</td>
<td>Application of indicators to show ICD-10 based data is sufficient.</td>
<td>By implication</td>
<td></td>
</tr>
<tr>
<td>8. Reporting burden</td>
<td>Application of indicators.</td>
<td>By implication</td>
<td></td>
</tr>
<tr>
<td>9. External and ecological validity</td>
<td>Explicit statement of assumptions used.</td>
<td>5.4.1</td>
<td></td>
</tr>
</tbody>
</table>

5.4.6. Analytical packages

All regression analyses were conducted using SAS version 9.2 using either the PROC LOGISTIC or PROC GLIMMIX procedures.
5.5. Results

5.5.1. Importance

5.5.1.1. National-level results

Across 147 acute English NHS hospitals, 91,936 stroke admissions were identified in the period April 2009 to March 2010. Of these, 2,522 (2.7%) died on the same day as admission, 15,846 (17.2%) died within 30 days of admission and 19,721 (21.5%) died before discharge. Of those patients meeting the inclusion criteria (discussed in Figure 11), 69.7% were scanned within one day of admission, 2.6% received thrombolysis, 5.3% had aspiration pneumonia, 72.8% were discharged to their normal place of residence, and 11.0% were readmitted as an emergency within 30 days of discharge.

5.5.1.2. Amenable

The wide range in performance suggests scope for improvement. For instance, the rate of scanning by-next-day ranged from 41.5% to 86.9% and for 30-day in-hospital mortality from 10.1% to 23.1%. Similar work I conducted as part of a team using the scanning indicators is reported in the discussion (paragraph 5.6.2).

5.5.2. Precision

Displaying the hospital-level data on funnel plots highlighted the variation in performance (Figure 12). With the exception of the measure for emergency readmissions, all the indicators identified at least one hospital as having performance outside the 99.8% control limits (Table 24). The number of stroke admissions per hospital ranged from 171 to 1532.
PART II: Stroke Care

Figure 12: Funnel plots of hospital-level performance across six indicators spanning the hospital stroke care pathway for 2009-10

a. Same day scanning

b. Thrombolysis rates

c. Aspiration pneumonia rates

d. 30-day in-hospital mortality rates

e. Discharge to usual place of residence within 56 days

f. 30-day emergency readmission rates

Note: Each dot represents a hospital. The horizontal line refers to national average; short-gauge dotted line refers to p<0.025 significance level; long-gauge dotted line refers to p<0.001 significance level.
Table 24: Number of trusts identified as having statistically significant above or below average performance

<table>
<thead>
<tr>
<th>Label</th>
<th>Indicator</th>
<th>National rate</th>
<th>Range</th>
<th>Number of hospitals higher than average</th>
<th>Number of hospitals lower than average</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>147 hospitals</td>
<td>[min, max]</td>
<td>p&lt;0.001</td>
<td>p&lt;0.025</td>
</tr>
<tr>
<td>a</td>
<td>Same day scan</td>
<td>47.1%</td>
<td>[20.4% - 79.3%]</td>
<td>38</td>
<td>47</td>
</tr>
<tr>
<td>a.1</td>
<td>By-next-day scan</td>
<td>69.7%</td>
<td>[41.5% - 86.9%]</td>
<td>13</td>
<td>32</td>
</tr>
<tr>
<td>a.2</td>
<td>Combined scan¹</td>
<td>n/a</td>
<td>n/a</td>
<td>9</td>
<td>23</td>
</tr>
<tr>
<td>b</td>
<td>Thrombolysis</td>
<td>2.6%</td>
<td>[0% - 16.8%]</td>
<td>13</td>
<td>24</td>
</tr>
<tr>
<td>c</td>
<td>Aspiration pneumonia</td>
<td>5.3%</td>
<td>[1.6% - 12.5%]</td>
<td>10</td>
<td>21</td>
</tr>
<tr>
<td>d</td>
<td>30-day in-hospital mortality</td>
<td>17.2%</td>
<td>[10.1% - 23.1%]</td>
<td>0</td>
<td>8</td>
</tr>
<tr>
<td>e</td>
<td>Discharge to usual place of residence within 56 days</td>
<td>72.8%</td>
<td>[54.9% - 85.4%]</td>
<td>0</td>
<td>7</td>
</tr>
<tr>
<td>f</td>
<td>30-day emergency readmissions (all cause)</td>
<td>11.0%</td>
<td>[6.1% - 17.3%]</td>
<td>0</td>
<td>4</td>
</tr>
</tbody>
</table>

Note 1: Combined scan (a.2) was calculated as p<0.001 if both scan indicators had this value, and p<0.025 if both indicators at least p<0.025.

Note 2: Better care is indicated by higher than average performance for measures a, b, and e, and lower than average performance in measures c, d and f.

5.5.3. Construct validity

Across the indicators, there were six pairs of indicators which had a statistically significant correlation at the 95% level, of which two were significant at the 99.8% level (Table 25).
PART II: Stroke Care

Table 25: Coefficient of correlation between pairs of indicators

<table>
<thead>
<tr>
<th></th>
<th>By-next-day scan</th>
<th>Thrombolysis</th>
<th>Aspiration pneumonia</th>
<th>30-day in-hospital mortality</th>
<th>Discharge to usual place of residence within 56 days</th>
<th>30-day emergency readmissions (all cause)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Same day scan</td>
<td>0.570**</td>
<td>0.040*</td>
<td>0.035*</td>
<td>-0.040*</td>
<td>-0.023</td>
<td>0.004</td>
</tr>
<tr>
<td>By-next-day scan</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Thrombolysis</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Aspiration pneumonia</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>30-day in-hospital mortality</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Discharge to usual place of residence within 56 days</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note: Correlation significant at the 95% level marked with ‘*’; those at 99.8% level with ‘**’. Coefficient is Pearson’s r.

5.5.4. Minimum bias: coding practice

5.5.4.1. Coding depth

At a hospital level, the average number of distinct diagnosis codes used per admission ranges from 5.0 to 10.7. There was a statistically significant but weak correlation (p = 0.002, R² = 0.067) between coding depth (measured as average number of unique diagnosis codes in an admission) and performance against the aspiration pneumonia measure. Of the 25 hospitals identified at the 99.8% level in the original regression, 20 (80.0%) were again flagged as outliers at this significance level when coding practice was included in the regression, with no hospitals flagged as having statistically different performance highlighted in the reverse performance category.

5.5.4.2. Case ascertainment: episodes

Secondly, across hospitals the proportion of strokes diagnosed as ICD-10 code I64 – Stroke Unspecified varied from 0.2 to 42.6%; however, there is only a weak correlation (p = 0.125, R² = 0.016) between use of this code and performance in the outcome measure of in-hospital mortality within 30 days. There was a statistically significant but weak correlation (p = 0.04, R² = 0.028) between the proportion of patients without a specific stroke diagnosis and the hospital’s one-day scanning rates; this association was expected as a scan is required to determine whether a stroke is ischaemic or haemorrhagic.
5.5.4.3. Case ascertainment: date

Nationally, the date-of-stroke calculations (start date of spell with stroke diagnosed, as described in Figure 11 and in the methodology above) were the same for 97.4% (and within one day for 97.9%) of patients, as compared with using the superspell start date. In comparison, when using the date of the first stroke episode as the alternative, the corresponding figures were 88.7% (93.5%). At a hospital level, there was a statistically significant but weak linear correlation ($p = 0.006, R^2 = 0.051$) between the proportion of patients identified as having uncertain stroke date and performance in this measure. Of the 76 hospitals identified at the 99.8% level in the original regression, 50 (65.8%) were again flagged as outliers at this significance level when coding practice was included in the regression, with one further hospital (1.3%) flagged as having statistically high rather than low scan rates.

5.5.5. Risk model

The case-mix models were better at predicting 30-day in-hospital mortality ($c = 0.735, r^2 = 0.158$) than readmissions ($c = 0.622, r^2 = 0.039$).

Table 26: Model fit statistics

<table>
<thead>
<tr>
<th>Label</th>
<th>Indicator</th>
<th>Model fit statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>Same day scan</td>
<td>c = 0.616</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$r^2 = 0.062$</td>
</tr>
<tr>
<td>a.1</td>
<td>By-next-day scan</td>
<td>c = 0.645</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$r^2 = 0.100$</td>
</tr>
<tr>
<td>b</td>
<td>Thrombolysis</td>
<td>c = 0.771</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$r^2 = 0.108$</td>
</tr>
<tr>
<td>c</td>
<td>Aspiration pneumonia</td>
<td>c = 0.661</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$r^2 = 0.049$</td>
</tr>
<tr>
<td>d</td>
<td>30-day in-hospital mortality</td>
<td>c = 0.735</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$r^2 = 0.158$</td>
</tr>
<tr>
<td>e</td>
<td>Discharge to usual place of residence within 56 days</td>
<td>c = 0.664</td>
</tr>
<tr>
<td></td>
<td>(all cause)</td>
<td>$r^2 = 0.095$</td>
</tr>
<tr>
<td>f</td>
<td>30-day emergency readmissions</td>
<td>c = 0.622</td>
</tr>
<tr>
<td></td>
<td>(all cause)</td>
<td>$r^2 = 0.039$</td>
</tr>
</tbody>
</table>

Note: $r^2$: the max-rescaled r-square score reported from PROC LOGISTIC

5.6. Discussion

5.6.1. Summary of findings

The results show the potential for using hospital administrative data, meeting the Government’s intention to measure stroke care and, moreover, to highlight potentially significant variations in the quality and safety across the care pathway. Six measures of quality and safety, covering the acute stroke care pathway, were applied to English hospital administrative data identifying 91,936 strokes. Five of the six indicators identified hospitals with statistically outlying performance at the 99.8%
level. In particular, the analysis reiterated the high impact of the subject matter, with large numbers of potential adverse events.

5.6.2. Amenable: scope for improvement

For most of the measures there is no clinical consensus or guidelines for what actual levels are acceptable and, in these cases, hospitals should strive to operate at the level of the best performers. Indeed, regulators and guideline-setters could use this analysis to suggest levels of performance that are attainable. The exception to this rule is in access to a scan, for which extant guidelines recommend that all stroke patients should receive brain imaging “within a maximum of 24 hours after onset of symptoms” and, as such, the actual performance of 69.7% of patients receiving a scan within one day of admissions is unacceptable. While some of this shortfall in performance against the measure could be due to incomplete coding of the procedure, audit data in England suggest a similar figure. The level of performance also falls short of the Danish NIP proposed a benchmark that providers should give at least Proportion of patients who have a CT/MR scan on the day of admission. This project also suggested a standard of a maximum of 15% of stroke patients dying within 30 days which, again, was not met in England where 17.2% died within this timeframe.

Work I conducted as part of a team during the development of the indicator set and investigated the inequalities on this scanning indicator further since being a process measure and with only limited and explicit contraindications for the procedure, the expectation is for limited variation in a good performance health service. This analysis, again on English HES data but using a different time period (2008/09) showed that young patients were more likely to be scanned quickly than old patients (Adjusted OR for 10 year increase of age = 0.90; P<0.001; 95% CI = 0.89–0.91); the relations for comorbidity were also similar. These findings again support the case that improvements can be made.

5.6.3. Precision

By using funnel plots and this control limit, I was able to account for random variation. This technique was more appropriate than one stroke study which identified outliers as being those 10% from expected and, therefore, not accounting for the scale of random variation being linked to the number of cases. If the deviation were entirely due to random variation, you would only expect to identify one outlier for every 3 to 4 measures (given there are 147 hospitals in the sample). Further to chance, differences in performance may be due to case mix, how the data were collected, or quality of care. Given the ambition was to compare hospitals on the basis of the last of these factors, the previous two also need to be accounted for.
5.6.4. Minimum bias

5.6.4.1. Adjusting for case-mix

Previous work has highlighted the importance of adjusting for case mix when comparing performance in stroke care. I accounted for case mix using patient-level logistic regression to calculate expected number of events for the outcome measures. However, at a patient level, some significant case-mix factors for stroke, such as severity of stroke and pre-stroke function, are not directly recorded within the data and therefore some of the variation may still be caused by differences in case-mix. If we hypothesis that there are small areas of extreme case-mix but, at a larger geographical level, case-mix is more homogenous coupled with units and trusts predominantly treating patients from their local area then it holds that any bias from this potential ‘categorisation error’ will be further diminished by the large number of stroke patients admitted at each hospital, ranging from 171 to 1532.

I report some model fit statistics which suggest that the case-mix models were better at predicting mortality ($r^2 = 0.158$) than readmissions ($r^2 = 0.039$) or pneumonia ($r^2 = 0.049$). This information is not for making conclusion on the appropriateness of the case-mix adjustment – since it is not possible to disaggregated the effect of unmeasured differences in case-mix to differences in quality of care provided – but should be considered when interpreting the results.

One specific issue relating to case mix originates from stroke care being increasingly delivered in regional networks, whereby certain hospitals are responsible for the urgent care of patients, whilst the other hospitals may take responsibility for rehabilitation. In these cases, the ambulance service has bypass protocols so that patients requiring urgent treatment are taken directly to the designated hospital. Therefore, the hospital designated to provide urgent care is more likely to receive a higher percentage of patients eligible for thrombolysis and, as such, their thrombolysis rates will be higher. In contrast, hospitals that are not designated to provide urgent care receive fewer eligible patients and, as a result, their performance against the measures of urgent stroke care are likely to be worse.

5.6.4.2. Accounting for variation in coding practice

The variation in performance due to how data were collected is harder to disaggregate; some of the possible variations in coding practice, for example, would mimic recognised alternative hospital pathways. However, I showed that three coding issues that were central to the assumptions for extracting the data explained only a small proportion of the differences in performance against the measures.
5.6.5. Errors in underlying data

There are few recent studies specifically investigating the accuracy of coding stroke care in healthcare systems using the ICD-10 framework; however, one article on coding in England found that coding of stroke diagnoses was excellent and elsewhere administrative data have been recommended for use in tracking progress and identifying problems for further review.\(^{340,325}\)

Previous studies have suggested coding is improving and, specifically in the instance of stroke care, the fact that some of the codes are new - for example the scanning procedure codes were only introduced in 2006 - means that it can be expected that there will initially be under-use (and so under-reporting) against these codes and that recording will improve over time.\(^{341}\)

The analysis of coding practice highlighted variations in the data recording practice of hospitals, even where guidance exists. For example, the current ICD-10 Clinical Coding Instruction Manual “directs the coder that on emergency admissions for strokes it is of paramount importance that the coder assigns the code for stroke in the primary position”.\(^{333}\) However, variations in the number of stroke codes recorded in secondary diagnosis fields above what would be expected from different prevalence of stroke as a co-morbidity suggest there are, from hospital to hospital, differing proportions of stroke recorded in secondary diagnosis fields.

An English study evaluating the accuracy of the coding of stroke diagnosis found that the cause of errors was predictable, with confusion in coding of different types of strokes. This implies that further guidance and training may be needed to ensure consistency in coding.\(^{340}\) One recent suggestion to facilitate an improvement is to develop better relationships between coders and clinicians.\(^{306}\) This paper explicitly outlines the assumptions used for identifying strokes in HES data and, therefore, can form the basis of a debate within the medical and coding practitioner community about how to further develop these indicators and the dataset itself.

5.6.6. Face and content validity

Whilst previous studies have used hospital administrative data to measure the performance of aspects of stroke care, none have brought together the multiple facets of the care pathway and instead mostly focused on one area, such as mortality, and on national trends rather than hospital comparisons. A combination of process and outcome measures was used in this study, so benefiting from the advantages of process measures (which tend to be more sensitive to differences in the quality of care and offer a clear action for improvement) and outcome measures (greater intrinsic interest, high face validity, and can reflect all aspects of care, including those that are otherwise difficult to measure such as technical expertise and operator skill).\(^{337,342}\) Face validity can be derived from the fact that the indicators appeared in existing literature or clinical guidelines.
As described in the previous chapter, overall level of quality of care at a provider-level cannot be deduced from the results of only a small set of indicators; however, two-thirds of studies within the literature review applied only one indicator. This study included six indicators covering the acute care pathway and, therefore, is likely to have better content validity (extent to which a measure represents all facets of the subject being investigated) over these previous more limited evaluations.

### 5.6.7. Construct validity

In the absence of a gold standard dataset to compare results, some assurance on validity can be taken by comparison with other indicators of performance. I investigated whether there were associations between hospitals’ performance across the different measures. In the existing literature, one study compared performance across different indicators constructed from the same dataset; however, in this instance, few pairs had significant correlation. One would expect some correlation between good performance across all the indicators with, for example, good management of a stroke unit likely to affect all the quality and safety indicators to some extent. I also predicted that there were two specific clinical explanations, with correlations expected if either:

1. two indicators measure different events on a defined clinical pathway; or
2. an outcome indicator measures the results of a process indicators.

Indeed these two categories could explain the significant correlations that I identified between: same day scanning and next day scanning (positive correlation); same day scanning and thrombolysis (positive); scanning rates and mortality (negative); and pneumonia and discharge to usual place of residence within 56 days (negative). However, the causation in this latter correlation, between a proxy outcome measure for swallow assessments and discharge destination, is unclear as developing pneumonia is likely to increase length of stay (and so decrease likelihood of discharge within 56 days) whilst a delayed discharge would inherently increase the chance of developing pneumonia in hospital. Further analysis reveals that at least a third (32.1%) of these pneumonia were diagnosed within 4 days of stroke; however, since the date of diagnosis is not specifically recorded in HES data this figure represents an absolute minimum and, as such, I cannot infer any causal trends between length of stay and pneumonia rates.

The one unexpected, statistically significant result was the positive correlation between same day scanning and aspiration pneumonia rates. I expected that hospitals with high scanning rates would also be likely to have good processes to ensure urgent swallow assessments and, therefore, have lower rates of aspiration pneumonia (so a negative correlation). A plausible explanation for this unexpected result is that a hospital with more comprehensive coding practices is more likely to
record both the scanning procedure code and pneumonia diagnosis code compared to a hospital with less rigorous coding.

5.6.8. Strengths of this study to other stroke measurement efforts

In the NSSA, there is already a comprehensive tool for measuring hospital performance in stroke care. However, the underlying data are self-reported, and the NSSA’s two-year reporting cycle coupled with time-lag before publication prohibits its use as a real-time monitoring tool. A recent positive development has been the introduction of the Stroke Improvement National Audit Programme (SINAP) which collects real-time data on stroke patients albeit this only covers the first three days of care and does not include all hospitals. The relative advantages of using administrative data were established earlier (paragraph 2.2, p42), but recall that HES has the advantage of being: longitudinal; timely; covering all hospital admissions; and being relatively cheap, costing £1 per record to collect compared with around £10 - £60 per record for clinical registers.\textsuperscript{89} Whilst NCHOD have made use of some of the potential advantages of administrative data, in assessing a limited number of aspects of stroke care, their indicators are only updated annually and at the time of writing, hospital mortality from stroke was two years out of date.

5.6.9. Criterion validity

As well as these considerations on whether the indicators’ algorithms are functioning as intended, further assurance needs to be taken about whether the results of the indicators reflect actual performance. The most common process for validating the results is to compare against a gold standard. The most applied and comprehensive method for assessing hospital stroke care in England is the NSSA, which suggests it would be the most appropriate dataset to compare results based on administrative data. However, direct comparison with the published NSSA results is not easily achieved due to differences in definitions. In addition there is more statistical uncertainty associated with the NSSA clinical data, given that it is based on a small sample of around 60 patients per hospital.

5.6.10. Conclusion

This study shows that hospital administrative data have the potential for benchmarking the standard of hospitals performance over the acute stroke care pathway, in a timely way and without increasing the bureaucratic burden on hospitals. The additional analysis described in the discussion - on how scanning levels were dependent on patient characteristics - also support the argument that things could be improved.
Whilst neither the lack of credible gold standard nor evidence of bias from coding practice render the indicators redundant, it does temper the extent to which they could be used, such as for regulating hospitals. This study shows that HES provides the facility to record some key process and outcome measures across the care pathways in a cheap and timely manner. The chapter gives a comprehensive overview of feasibility and this should determine future application of these indicators. On the more sophisticated end of the scale, these results could be linked to structural measures, such as existence of specialist stroke units or a 24-hours service, to investigate their effectiveness (and also further evaluate the extent to which performance is amenable). With suitable modification and further development, these data could also be used routinely as part of the National Stroke Sentinel Audit and replace some aspects of the self-assessment.
Chapter 6.

Stroke temporal trends (longitudinal)
Overview

Context
The literature review revealed that less than a third of studies on stroke care had used the longitudinal nature of administrative data. This chapter investigates potential longitudinal applications of indicators to address both the gaps in the literature and also to provide evidence on the whether performance is amenable to how services are structured.

Methods
Application of indicators covering the acute stroke care pathway to investigate changes in performance over time, at both a national and regional level. Where appropriate, results were adjusted for case-mix using logistic regression and significant changes in performance were identified using control charts. Specifically, interrupted time series analysis with a concurrent control group study design was used in the regional analysis to assess the impact of the Hyper Acute Stroke Units (HASUs) in London.

Finding
The longitudinal analyses highlighted improvements in performance and a seasonal effect for some measures. Compared with areas outside London, 7 day in-hospital deaths rate reduced significantly following the restructuring of services, as did aspiration pneumonia. However, same day brain scans showed a small but significant reduction following the intervention, as well as a slowing down in the rate of increase.

What this chapter adds
The first discussion of how using temporal analyses can overcome commonly cited concerns on the application of administrative data and can be used to evaluate the impact of healthcare initiatives.

Acknowledgements and related papers
I undertook the national longitudinal analysis. The regional study was led by Roxanna Alexandra with all authors (listed below) involved in the study design, analysis and manuscript revision of the regional analysis.

6.1. Background

6.1.1. National trends
The review of stroke literature (Chapter 4) highlighted the unmet potential to use the longitudinal nature of administrative data to highlight potential quality and safety issues within stroke care. Measuring performance over time – rather than a snapshot cross-sectional analysis – might also mitigate against some of the hospital-level variations in unmeasured case-mix and coding practice. This was underlined by the hospital-level cross-sectional analysis (Chapter 5) which was found to be, at least in part, biased by variations in hospital-level coding practice. In England, previous studies have suggested that stroke care has improved but the trend in many aspects of stroke treatment and outcomes remains unclear.\textsuperscript{151,343}

6.1.2. Regional analysis
Such trend analysis is particularly useful when evaluating any change in structure or process of a particular provider unit, such as the implementation of regional stroke units, and this is investigated by presenting results of the evaluation of a regional reconfiguration. Prior to 2010, provision of stroke care in London was complex, with only 53% of patients treated on a dedicated stroke ward and care spread across a number of units.\textsuperscript{344} To improve the quality of service, eight Hyper Acute Stroke Units (HASUs) were established in London from February 2010. The units, which are dedicated to treating stroke patients, are open 24 hours, seven days a week to offer immediate access to stroke investigations and imaging, including CT brain scan and clot-busting thrombolysis drugs. These units replaced the previous disparate services for acute stroke care.

6.2. Objectives
In this chapter I investigate the potential to use the longitudinal nature of administrative data to identify important shortcomings in quality and safety in a way that overcomes some of the known limitations to such indicators. The regional analysis was conducted as a case example to show that the indicators can be used to assess the impact of interventions, in this case the HASU policy in London, and in doing so show that quality and safety can be influenced by the way services are provided. This part of the study is not intended as a full evaluation of the HASU policy but to demonstrate the potential of the indicator set I developed.
6.3. Methods

6.3.1. Indicators
The indicators used for this part of the study are similar to those used in the previous chapter (paragraph 5.3.1.2, p118) and were identified in the literature review (Chapter 4). The only amendment made to these indicators, for the purpose of this weekend part of the study, was to use 7-day rather than 30-day mortality. The indicators include both process and outcome measures:

a) Receiving a brain scan on the day of admission;
b) Receiving thrombolysis treatment;
c) Diagnosed with aspiration pneumonia (complication) in hospital;
d) 7-day in-hospital mortality;
e) Discharge to usual place of residence within 56 days; and
f) 30-day emergency readmission (all cause).

6.3.2. Analysis

6.3.2.1. Data extract
For the national trends, the details of stroke admissions from 1 April 2007 to 31 December 2009 were extracted from the Hospital Episode Statistics (HES) database. The indicator definitions were applied to the extract to obtain denominators and numerators, categorised for month of admission for stroke. For the regional analysis, data from 1 April 2006 to 31 March 2012 were extracted, to account for the differing demands on this analysis and the availability of additional data at the point that this analysis was conducted.

6.3.2.2. Risk-adjustment
Variables expected to influence the outcome of the association between the quality and safety indicators and day of admission were extracted: age; sex; socio-economic deprivation quintile; number of previous admissions; co-morbidities - Charlson index with weights derived from all England admissions; ethnic group; source of admission; and stroke type. The method for risk-adjustment is covered in paragraph 3.6.2, p79.

6.3.2.3. National data analysis
National performance across the six measures was produced, disaggregated by month and, where appropriate, multiple linear regressions were used to adjust to the effect of the covariates. Results
were also displayed by plotting – using April 2007 as a reference – odds ratios and 95% confidence intervals, by month of admission for stroke.

6.3.2.4. Regional data analysis
For the regional analysis, the quarterly unadjusted rates for the process and outcome indicator before and after the implementation of HASU policy were plotted. Using an interrupted time series (ITS) negative binomial regression model, the impact of HASU policy was assessed. While taking the underlying trend into account, this model estimates both the level and slope changes of the time trend after the policy change. A dummy variable (coded as 0 for the period before and 1 after implementation of the policy) and a continuous variable (set to 0 prior HASU policy, then equal to the length of time since HASU policy was implemented) were included. As we use quarterly data, the beginning of the reorganisation period is modelled as January instead of February (and ending in June).

To remove potential effect of other national wide policies on our outcomes, a comparison group was included in the ITS model using data from rest of England. To do so, a dummy variable representing region (coded as 1 for London and 0 for other areas in England) was included, and the model was fitted with interaction terms between the dummy variable for region and post-intervention step/slope terms. As the model estimates the effect of HASU after adjusting for a concurrent comparison group, a negative/positive change in the step (or slope) will indicate a larger decrease/increase in the rates in the intervention group (London) compared with the control group (the rest of England).

6.3.2.5. Software
Statistical analyses were conducted using SAS version 9.2 software’s PROC GENMOD and PROC LOGISTIC procedure for regression analysis.

6.4. Results

6.4.1. National

6.4.1.1. Overview of results
The number of stroke admissions rose by 9.8% from 85,287 in 2007/08 to 93,621 in 2009/10 (Table 27). Between these two years, there were reported improvements in scanning (from 29.2% in 2007/08 to 46.5% in 2009/10), thrombolysis (0.3 to 2.6%), 30-day in-hospital mortality (20.1 to 17.1%), and discharge to usual place of residence (68.7 to 72.6%). Conversely, reported performance
got worse for aspiration pneumonia (4.8 to 5.2%) and 30-day all-cause emergency readmissions (9.2 to 11.05).

### Table 27: Annual trend in performance

<table>
<thead>
<tr>
<th>Label</th>
<th>Indicator</th>
<th>2007/08</th>
<th>2008/09</th>
<th>2009/10 ^</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>n = 85,287</td>
<td>n = 89,712</td>
<td>n = 93,621</td>
</tr>
<tr>
<td>a</td>
<td>Same day scan</td>
<td>29.2</td>
<td>37.4</td>
<td>46.5</td>
</tr>
<tr>
<td>a.1</td>
<td>By-next-day scan</td>
<td>50.4</td>
<td>60.0</td>
<td>68.7</td>
</tr>
<tr>
<td>b</td>
<td>Thrombolysis</td>
<td>0.3</td>
<td>1.1</td>
<td>2.6</td>
</tr>
<tr>
<td>c</td>
<td>Aspiration pneumonia</td>
<td>4.8</td>
<td>5.0</td>
<td>5.2</td>
</tr>
<tr>
<td>d</td>
<td>30-day in-hospital mortality</td>
<td>20.1</td>
<td>18.8</td>
<td>17.1</td>
</tr>
<tr>
<td>e</td>
<td>Discharge to usual place of residence within 56 days</td>
<td>68.7</td>
<td>70.2</td>
<td>72.6</td>
</tr>
<tr>
<td>f</td>
<td>30-day emergency readmissions (all cause)</td>
<td>9.2</td>
<td>10.6</td>
<td>11.0</td>
</tr>
</tbody>
</table>

**Note:** ^ Proportions for 2009/10 may differ slightly from those reported in the previous chapter, where the results were limited to only those patients admitted to recognised acute trusts.

#### 6.4.1.2. National longitudinal trends

Disaggregating national performance by month suggests there are different types of longitudinal trends (Figure 13):

1. **Steady increases.** There has been a steady, statistically significant increase in scanning rates (panel a) over time, which can be closely approximated by a linear regression (R-squared 0.96, $\beta = +0.06, p < 0.001$). Similarly, thrombolysis rates have increased gradually over time (panel b).

2. **Overall trend with seasonal effect.** Mortality rates have also improved over time (R-squared 0.28, $\beta = -0.006, p=0.002$) but there is also evidence of a substantial seasonal effect (panel d) with rates highest around December. The pattern is similar for the rate of patients being discharged to usual place of residence (panel e) where there has been a general improvement with a seasonal increase around July.

3. **No clear trend.** Figures for pneumonia (panel c) and readmission (panel f) rates showed only minimal seasonal or overall longitudinal trend.
PART II: Stroke Care

Figure 13. Performance by month¹

a. Immediate scan (unadjusted)

b. Thrombolysis (unadjusted)²

c. Pneumonia (adjusted)

d. Death within 30 days (adjusted)

e. Discharge to usual place of residence (adjusted)

f. Readmission (adjusted)³

Note: 1. Data points represent odds ratios, with April 2007 used as a reference (1.00); vertical ranges, 95% confidence intervals.

2. The low number/rate for the reference day on the Thrombolysis chart results in large OR for later months; this is discussed later.

3. For the indicator of 30-day emergency readmissions, data for Jan-08 to Mar-08 has been removed due to data issues.
6.4.2. Regional trends

The regional analysis included 536,034 stroke admissions in England over the 6-year period to March 2012. Around one-in-nine (11.5%) of these admissions occurred in London. The average annual admission rate in London was 42,375 per year prior to the intervention. This admission rate increased slightly (by 4%) post intervention in London but remained static in the rest of England. The national rates are given in Table 28. Over the entire study period, London performed better than the rest of England on scanning, thrombolysis, and mortality rates, with the inverse true for pneumonia, discharge of usual place of residence, and emergency readmission rates.

<table>
<thead>
<tr>
<th>Table 28. Summary of differences in performance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Indicator</td>
</tr>
<tr>
<td>Same day scanning</td>
</tr>
<tr>
<td>Thrombolysis</td>
</tr>
<tr>
<td>Aspiration pneumonia</td>
</tr>
<tr>
<td>7-day in-hospital mortality</td>
</tr>
<tr>
<td>Discharge to usual place of residence within 56 days</td>
</tr>
<tr>
<td>30-day emergency readmissions</td>
</tr>
</tbody>
</table>

Table 29 sets out the crude rates before and after the intervention for both London and the rest of England. In the these regions, there were improvements in scanning rates, thrombolysis, mortality and discharge destination. Conversely, readmission rates increased in both London and the rest of England. While crude rates of aspiration pneumonia decreased in London (6.3% to 5.6%) they increased elsewhere (4.9% to 5.4%).
Table 29. Crude rates before and after intervention

<table>
<thead>
<tr>
<th>Indicator</th>
<th>London, rate (number)</th>
<th>Rest of England, rate (number)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Before</td>
<td>After</td>
</tr>
<tr>
<td>n = 37 051</td>
<td>n = 19 270</td>
<td>n = 289 491</td>
</tr>
<tr>
<td>Same day scanning</td>
<td>38.3% (14 194)</td>
<td>62.5% (12 047)</td>
</tr>
<tr>
<td>Thrombolysis</td>
<td>0.7% (265)</td>
<td>4.9% (947)</td>
</tr>
<tr>
<td>Aspiration pneumonia</td>
<td>6.3% (2 326)</td>
<td>5.6% (1 088)</td>
</tr>
<tr>
<td>7-day in-hospital mortality</td>
<td>8.7% (3 235)</td>
<td>6.8% (1 311)</td>
</tr>
<tr>
<td>Discharge to usual place of residence within 56 days</td>
<td>55.9% (20 707)</td>
<td>59.9% (11 540)</td>
</tr>
<tr>
<td>30-day emergency readmissions</td>
<td>8.9% (3 282)</td>
<td>10.4% (2 006)</td>
</tr>
</tbody>
</table>

Notes: ‘Before’ denotes April 2006 to December and ‘After’ denotes July 2010 to March 2012

Table 30 shows the results of the interrupted time series analysis for London, relative to the rest of England. There was a significant step change fall in aspiration pneumonia (OR 0.81) and 7 day in-hospital death (0.93). Same day brain scans appeared to decrease relative to England (0.91). There was a decrease in slope relative to the rest of England (0.95), however there was also a reduction in slope for same-day brain scans (0.98) and discharge to usual place of residence within 56 day (0.99).
Table 30. Time-Series regression analysis of stroke indicators in London relative to the rest of England

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>OR</td>
<td>95% CI</td>
</tr>
<tr>
<td>Same day scanning</td>
<td>0.91</td>
<td>(0.88-0.94)</td>
</tr>
<tr>
<td>Thrombolysis</td>
<td>1.11</td>
<td>(0.95-1.29)</td>
</tr>
<tr>
<td>Aspiration pneumonia</td>
<td>0.81</td>
<td>(0.74-0.87)</td>
</tr>
<tr>
<td>7-day in-hospital mortality</td>
<td>0.93</td>
<td>(0.86-0.99)</td>
</tr>
<tr>
<td>Discharge to usual place of</td>
<td>0.96</td>
<td>(0.93-0.99)</td>
</tr>
<tr>
<td>residence within 56 days</td>
<td></td>
<td></td>
</tr>
<tr>
<td>30-day emergency readmissions</td>
<td>1.02</td>
<td>(0.96-1.09)</td>
</tr>
</tbody>
</table>

Note: Selected time periods based on pre-intervention trend linearity: Same-day brain scan (July 2007 – March 2012); Thrombolysis (October 2007 – March 2012); Aspiration pneumonia (April 2007 – March 2012); 7 day in-hospital death (April 2007 – March 2012); Discharge to usual place of residence 56 day (April 2008 – March 2012); 30 day emergency readmissions (April 2007 – March 2012). Transition period (January to June 2010) removed.

6.5. Discussion

6.5.1. Longitudinal

The longitudinal analysis showed that administrative data could be used to identify potentially important trends in stroke care quality and safety. In particular, the analysis showed significant improvements over four of the six indicators. Interestingly, the analysis also showed a strong seasonal trend in stroke outcomes of death rates and proportion of stroke patients discharged to their usual place of residence. This is an area which, to date, has been poorly researched.

A key strength of this on-going and longitudinal application of administrative data is that it is not possible with the prominent existing tools for monitoring stroke care in England; the NSSA is
collected over a short period of time and the Stroke Improvement National Audit Programme (SINAP) only covers the urgent part of the care pathway.

6.5.1. Regional analysis

The results of the interrupted time series analysis are consistent with the expected effect if the reorganisation of stroke services had improved the quality of care in London. Importantly, 7 day in-hospital deaths showed a significant step reduction following the restructuring of services, as did aspiration pneumonia when compared to trends in the rest of England. Interestingly, same day brain scans showed a small but significant reduction compared with the rest of England following the intervention, as well as a slowing down in the rate of increase. This may be due to a possible ceiling effect, i.e., a limit beyond which any improvement is less likely to continue. In any case, London scan rates are consistently higher than the rest of England.

The analysis supports the previously (paragraph 5.6.2, p130) made argument that performance against the stroke indicator set is amenable. A wider evaluation and discussion about the value of HASUs is beyond the scope of this project and, rather, this analysis was presented to support the evidence against the analytical framework item of ‘impact’.

6.5.2. Limitations

6.5.2.1. Changes in coding practice

This application mitigates against the bias from variations in coding practice between hospitals, as even if there is evidence of a change in distribution of patients between hospitals over time then this clustering can be accounted for in such regression analyses. As discussed earlier, the level of thrombolysis coding might be expected to increase due to financial incentives attached with the recording of this treatment and, also, due to the low stating point because thrombolysis has only recently been licensed in England.

6.5.2.2. Interpreting comparison over time

The analysis also highlighted the potential issues with making comparisons over time. In particular, the choice of using the first month as the reference was unhelpful when interpreting the change in thrombolysis performance since this first month was an extreme performance and likely to be a consequence of a lack of incentive at the time to code thrombolysis correctly and the treatment not being formally approved. Indeed, the odds ratio are substantially different depending on what reference month is chosen; for example, with reference to April 2007 (month 1) the odds ratios for month 2 (May 2007) and month 33 (December 2009) are 2.109 (95% confidence interval 0.680 – 6.539) and 33.670 (12.458 – 91.004), whereas if April 2008 (month 13) is used as reference then the
estimated odds ratios are more precise: 0.114 (0.049–0.268) and 2.846 (2.027–3.994), respectively.

The issues are not isolated to just thrombolysis nor national-level analysis. Statistical control charts are common techniques for monitoring performance in healthcare. Figure 14 shows performance for three anonymised hospitals (labelled Hospital A, B, C which are based on actual data from the study) against the same day scanning code. Taking the rule that a trend of six of more consecutive increases (or decreases) signifies unexpected variation, then Hospital A would be the only hospital identified as a potential ‘good performing’ outlier although would not necessarily be the obvious interpretation when comparing the trends and overall performance levels. The issue described is particularly problematic when changes in coding practice could mimic actual improvements (i.e. a hospital which is aiming to improve the recording of scans will automatically appear to be performing better).

Figure 14. Example of hospital trends

6.5.3. Conclusion

This chapter show that it is possible to address some of the potential bias from case-mix and coding practice using longitudinal analyses; however, given the changes in coding practice over time which coincides with potential changes in service provision (and therefore hard to detect) this application is still susceptible to bias. The regional analysis had the advantage of providing evidence that some of the shortcomings in care are amenable to health care organisation.
Chapter 7.
Stroke temporal trends (day of admission)
Chapter 7: Stroke temporal trends (weekends)

Overview

Context
The literature review revealed that no study had comprehensively assessed weekend care. This chapter investigates potential temporal applications of indicators to address both the gaps in the literature and also overcome the shortcomings in validity relating to cross-sectional application of these indicators as identified in the previous chapter.

Methods
Application of indicators covering the acute stroke care pathway to investigate the association between day of admission and adverse events. Where appropriate, results were adjusted for case-mix using logistic regression and significant changes in performance were identified using control charts.

Finding
The day-of-the-week analysis showed performance across five of the six measures was significantly lower at weekends at the 99.8% level. In particular, the rate of 7-day in-hospital mortality for Sunday admissions was 11.0% (adjusted odds ratio 1.26; 95% Confidence Interval 1.16 – 1.37 with Monday as reference) compared to an average of 8.9% for weekday admissions.

What this chapter adds
This chapter provides the most comprehensive evaluation of the “weekend effect” in stroke care.

Acknowledgements and related papers
7.1. Background

7.1.1. Measuring the ‘weekend effect’

Further to using the longitudinal nature of administrative data, an alternative temporal approach which could mitigate some of the known coding issues is to evaluate the association between quality of care and day of admission. Previous studies, across a range of countries, have identified higher mortality in patients admitted on weekends across a range of medical conditions; a phenomenon termed the ‘weekend effect’. This calls into question the idea that quality of care is equal irrespective of when you present at hospital. Similarly, there have also been a limited number of international studies investigating the effect specifically in stroke care that have suggested poorer access to treatments and worse outcomes at weekends, including increased mortality and fewer patients returning to their usual place of residence. However, other studies have not identified a significant association between day of admission and stroke mortality rates. Therefore, debate remains over the existence and extent of the weekend effect in stroke care. The studies on out-of-hours stroke care are limited in number and most of these fail to stand up to the criticism of primarily focusing on short-term mortality and therefore not capturing wider aspects of the quality and safety of care.

This chapter is unique in providing a comprehensive and current assessment of the degree to which quality and safety of stroke care is affected by whether a patient is admitted at the weekends, highlighting the potential for identifying clinically important issues using this readily available resource.

7.2. Objectives

In this chapter I investigate the potential to use the longitudinal nature of administrative data to identify important shortcomings in quality and safety in a way that overcomes some of the known limitations to such indicators.

7.3. Methods

7.3.1. Indicators

The indicators used for this part of the study are similar to those used in the previous chapter (paragraph 5.3.1.2, p118) and were identified in the literature review (Chapter 4). The only amendment made to these indicators, for the purpose of this weekend part of the study, was to use 7-day rather than 30-day mortality in keeping with the suggestion that it is a more appropriate
Chapter 7: Stroke temporal trends (weekends)

timeframe to evaluate the association between day of admission and mortality. The indicators include both process and outcome measures:

   a) Receiving a brain scan on the day of admission;
   b) Receiving thrombolysis treatment;
   c) Diagnosed with aspiration pneumonia (complication) in hospital;
   d) 30 (or 7)-day in-hospital mortality;
   e) Discharge to usual place of residence within 56 days; and
   f) 30-day emergency readmission (all cause).

7.3.2. Weekend analysis

7.3.2.1. Data extract
For the weekend study, stroke admissions from 1 April 2009 to 31 March 2010 only were used. The indicator definitions were applied to the extract to obtain denominators and numerators, categorised for day of admission for stroke. The weekend was defined as the period from midnight on Friday to midnight on Sunday, with all other times defined as weekdays (the time of admission is not captured in HES).

7.3.2.2. Data analysis
Variables expected to influence the outcome of the association between the quality and safety indicators and day of admission were extracted: age; sex; socio-economic deprivation quintile; number of previous admissions; co-morbidities - Charlson index with weights derived from all England admissions, ethnic group; source of admission; and stroke type. Where only one figure is given, process measures were reported as unadjusted rates (paragraph 5.4.2.2, p122). The method for risk-adjustment is covered in paragraph 3.6.2, p79.

Descriptive analysis of the patients, categorised by weekday or weekend admission for stroke were produced. Initially unadjusted (crude) rates for each of the six indicators were calculated for weekday and weekend admissions (these are plotted for the process measures: access to a scan and thrombolysis). Multiple logistic regression analyses were used to account for the covariates and estimate adjusted odds, across each indicator, for weekend compared with weekday admissions. Results were also displayed by plotting – using Monday as a reference – odds ratios and 95% confidence intervals, by day of the week of admission for stroke.
Analyses were carried out using SAS Version 9.2, using the PROC LOGISTIC procedure for regression analyses. I did not adjust for the clustering of patients within hospital as the hospital-level effects were found to be small.

### 7.3.2.3. Estimating impact

Using regression analysis on just weekday admissions, probabilities of in-hospital death and discharge to usual place of residence for different patient characteristics were derived. By matching these probabilities for each weekday admission, based on the patient’s characteristics, indirectly standardised estimates for the outcomes as if those weekend patients had had similar rates as their weekday counterparts were calculated, with these particular results reported in the discussion.

### 7.3.2.4. Consultant specialty

To investigate one of the possible organisational factors that could cause variations in the performance against the process and outcome measures, an additional indicator – *specialty of responsible consultant* – as a proxy for whether a patient is admitted to a designated stroke unit was calculated.

### 7.4. Results

#### 7.4.1.1. Descriptive overview

Across English NHS hospitals, I identified 93,621 stroke admissions in the period April 2009 to March 2010. Of these, 8,722 (9.3%) died in-hospital within 7 days of admission and 16,013 (17.1%) died within 30 days. Of those patients meeting the inclusion criteria for each indicator, 46.5% were scanned on the day of admission, 2.6% received thrombolysis, 5.2% had aspiration pneumonia, 72.6% were discharged to their normal place of residence, and 11.0% were readmitted within 30 days of discharge.
Of the stroke admissions, 24.9% (23 297) were admitted at weekends. The number of stroke patients admitted declined throughout the week from 15.6% on Monday to 12.4% on Sunday (Figure 15).

**Figure 15. Number of stroke admissions, by day of presentation**

![Bar chart showing number of stroke admissions by day of presentation.](image)

Table 31 gives a description of the characteristics of the patients included in the study.

**Table 31. Characteristics of the study population**

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Weekday % of admissions (number of cases)</th>
<th>Weekend % of admissions (number of cases)</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age of admission in years (s.d)</td>
<td>73.8 (14.8)</td>
<td>74.5 (14.6)</td>
<td>0.048</td>
</tr>
<tr>
<td>Sex</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>48.5 (34 096)</td>
<td>47.2 (10 990)</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Female</td>
<td>51.5 (36 228)</td>
<td>52.8 (12 307)</td>
<td></td>
</tr>
<tr>
<td>Index of multiple deprivation quintile</td>
<td></td>
<td></td>
<td>0.983</td>
</tr>
<tr>
<td>1 least deprived</td>
<td>16.8 (11 836)</td>
<td>16.9 (3 947)</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>20.6 (14 507)</td>
<td>20.7 (4 817)</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>21.4 (15 048)</td>
<td>21.3 (4 958)</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>20.9 (14 706)</td>
<td>21.1 (4 903)</td>
<td></td>
</tr>
<tr>
<td>5 most deprived</td>
<td>19.4 (13 668)</td>
<td>19.3 (4 491)</td>
<td></td>
</tr>
<tr>
<td>6 unclassified</td>
<td>0.8 (559)</td>
<td>0.8 (181)</td>
<td></td>
</tr>
<tr>
<td>Charlson index of comorbidity</td>
<td></td>
<td></td>
<td>0.100</td>
</tr>
<tr>
<td>0 no comorbidity</td>
<td>45.6 (32 034)</td>
<td>45.4 (10 573)</td>
<td></td>
</tr>
<tr>
<td>1-4*</td>
<td>13.8 (9 715)</td>
<td>14.0 (3 256)</td>
<td></td>
</tr>
</tbody>
</table>
### PART II: Stroke Care

<table>
<thead>
<tr>
<th>Group</th>
<th>Source of admission</th>
<th>Admission Type</th>
<th>Source of admission</th>
<th>Admission Type</th>
<th>Source of admission</th>
<th>Admission Type</th>
<th>Source of admission</th>
<th>Admission Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>5-9*</td>
<td>0</td>
<td>69.6 (48 960)</td>
<td>69.6 (48 960)</td>
<td>0</td>
<td>70.1 (16 327)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10-14*</td>
<td>1</td>
<td>19.3 (13 586)</td>
<td>19.3 (4 484)</td>
<td>1</td>
<td>19.3 (4 484)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15-24*</td>
<td>2</td>
<td>6.6 (4 650)</td>
<td>6.2 (1 444)</td>
<td>2</td>
<td>6.2 (1 444)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>25-50 highest comorbidity*</td>
<td>3</td>
<td>4.5 (3 128)</td>
<td>4.5 (1 042)</td>
<td>3</td>
<td>4.5 (1 042)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Number of admissions in previous 12 months**: 0.157

- **0**: 69.6 (48 960) 70.1 (16 327)
- **1**: 19.3 (13 586) 19.3 (4 484)
- **2**: 6.6 (4 650) 6.2 (1 444)
- **3**: 4.5 (3 128) 4.5 (1 042)

**Source of admission**: < 0.001

- **Admitted from home***: 89.2 (62 755) 89.9 (20 939)
- **Hospital Transfer***: 9.7 (6 799) 9.1 (2 115)
- **Other or Unknown***: 1.1 (770) 1.0 (243)

**Emergency**: 0.001

- **non-emergency admission**: 5.6 (3 925) 2.77 (645)
- **emergency admission**: 94.4 (66 399) 97.2 (22 652)

**Stroke Diagnosis**: 0.038

- **I60**: 5.1 (3 570) 5.0 (1 175)
- **I61**: 11.2 (7 864) 11.8 (2 738)
- **I62**: 6.2 (4 376) 5.8 (1 352)
- **I63**: 58.6 (41 188) 58.5 (13 628)
- **I64**: 19.0 (13 326) 18.9 (4 404)

**Total**: 75.1 (70324) 24.9 (23297)

**Notes**: $\chi^2$ test for association, p<0.001

* grouped for purposes of this table
### 7.4.1.2. Association between admission at weekends and performance

Table 32 shows the results of the association between weekday/weekend admission and performance in the six measures of quality and safety. There were statistically significant associations in five of the indicators (the exception being readmissions), all of which were consistent with a lower standard of care at weekends. The largest effects were seen in the lower rates of same-day scanning (unadjusted odds ratio [OR] 0.83; 95% Confidence Interval 0.81 – 0.86), thrombolysis (unadjusted OR 0.82; 0.73 – 0.92) and higher rates of in-hospital mortality (adjusted OR 1.18; 1.12 – 1.24) for weekend strokes.

**Table 32. Association between weekday/weekend admission and indicators of quality and safety of care**

<table>
<thead>
<tr>
<th>Indicator</th>
<th>Weekday admission</th>
<th>Weekend admission</th>
<th>p-value</th>
<th>OR (95% CI) (weekday as reference)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Same day scanning</td>
<td>47.6 (32521)</td>
<td>47.8</td>
<td>&lt;0.0001</td>
<td>0.833 (0.808 – 0.858)^</td>
</tr>
<tr>
<td>Thrombolysis</td>
<td>2.69 (1152)</td>
<td>2.70</td>
<td>0.001</td>
<td>0.819 (0.730 – 0.919)^</td>
</tr>
<tr>
<td>Aspiration pneumonia</td>
<td>5.04 (3545)</td>
<td>5.08</td>
<td>0.003</td>
<td>1.105 (1.035 – 1.180)</td>
</tr>
<tr>
<td>7-day in-hospital mortality</td>
<td>8.89 (6250)</td>
<td>8.98</td>
<td>&lt;0.0001</td>
<td>1.178 (1.120 – 1.240)</td>
</tr>
<tr>
<td>Discharge to usual place of residence within 56 days</td>
<td>73.0 (40692)</td>
<td>72.9</td>
<td>&lt;0.0001</td>
<td>0.918 (0.883 – 0.954)</td>
</tr>
<tr>
<td>30-day emergency readmissions within</td>
<td>11.1 (6165)</td>
<td>11.1</td>
<td>0.801</td>
<td>0.993 (0.940 – 1.049)</td>
</tr>
</tbody>
</table>

**Notes:**

^ For Unadjusted rates. The p-value and OR for the other measures are for adjusted rates

CI denotes confidence interval.
7.4.1.3. **Association between day of admission and performance**

Figure 16 shows the odds ratios by day of admission, with Monday as a reference, for the six indicators. For the measure of same-day scanning, the highest performing day was Friday with 49.8% of patients accessing a scan (unadjusted OR 1.16; 1.11 – 1.22 with Monday as the reference) whereas Sunday had the lowest rate at 42.0% (unadjusted OR 0.80; 0.88 – 0.97). For thrombolysis, the best performing day was Wednesday when 2.9% of stroke patients received the treatment (unadjusted OR 1.16; 0.98 – 1.37) with Sunday the worst at 2.2% (unadjusted OR 0.82; 0.68 – 0.99), however this latter result is no longer statistically significant when adjusting for case mix (adjusted OR 0.84; 0.70 – 1.02).

Rates of aspiration pneumonia were higher on Saturdays and Sundays (both 5.7%) than the Monday reference (5.0%); however this was not significant at the 5% level with p-values of 0.12 and 0.05 respectively. The highest rate of 7-day in-hospital mortality was for patients admitted for stroke on Sundays (adjusted OR 1.26; 1.16 – 1.37) at 11.0% compared with an average for weekday admissions of 8.9%. Both Saturday (adjusted OR 0.93; 0.87 – 0.99) and Sunday (adjusted OR 0.93; 0.87 – 0.99) had lower rates of discharge to usual place of residence within 56 days at 71.4% and 71.2% respectively, with an average for weekday admission of 73.0%. The association between day of admission and emergency readmission rates were not significant (p-value 0.60).
Figure 16. Performance by day of admission

a. Immediate scan (unadjusted)

b. Thrombolysis (unadjusted)

c. Pneumonia (adjusted)

d. Death within 7 days (adjusted)

e. Discharge to usual place (adjusted)

f. Readmission (adjusted)
7.4.1.4. **Age stratified results**

There was also some evidence that the weekend effect might be more pronounced for certain patient groups, with a more prominent effect on 7-day in-hospital mortality for patients aged 0 to 44 (adjusted OR 1.61; 1.23 – 2.12) than for patients aged 85 and over (1.13; 1.03 – 1.23) (Table 33).

**Table 33. The variation in ‘weekend effect’ across different age groups, for 7-day in-hospital mortality**

<table>
<thead>
<tr>
<th>Age group</th>
<th>Weekday unadjusted rate, % (admissions)</th>
<th>Weekend</th>
<th>Odds ratio v. weekday rate, adjusted (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 to 44</td>
<td>5.40 (3 221)</td>
<td>8.82 (986)</td>
<td>1.612 (1.225 – 2.123)</td>
</tr>
<tr>
<td>45 to 54</td>
<td>6.17 (4 523)</td>
<td>9.26 (1 414)</td>
<td>1.463 (1.170 – 1.829)</td>
</tr>
<tr>
<td>55 to 64</td>
<td>6.47 (8 344)</td>
<td>7.36 (2 527)</td>
<td>1.097 (0.918 – 1.311)</td>
</tr>
<tr>
<td>65 to 74</td>
<td>7.25 (14 075)</td>
<td>8.30 (4 565)</td>
<td>1.111 (0.978 – 1.261)</td>
</tr>
<tr>
<td>75 to 84</td>
<td>9.25 (22 809)</td>
<td>10.89 (7 660)</td>
<td>1.209 (1.108 – 1.320)</td>
</tr>
<tr>
<td>Over 85</td>
<td>12.26 (17 352)</td>
<td>13.91 (6 145)</td>
<td>1.128 (1.033 – 1.232)</td>
</tr>
</tbody>
</table>

Notes: CI denotes confidence interval.

7.4.1.5. **Association between consultant specialty and performance**

The most common specialty of consultants responsible for the first episode of care in hospital following a patient’s stroke was ‘general medicine’ (48.1%) followed by ‘geriatric medicine’ (26.3%). Table 34 sets out the difference in specialty of responsible consultants between weekday and weekend strokes, with the latter slightly more likely to be under the responsibility of a general medic or neurosurgeon, and less likely under geriatric medics. Aggregating the specialties into two groups showed that weekend strokes are less likely to be initially under the responsibility of neuro- or geriatric specialists (33.0% with general medics, A&E medics and other specialties accounting to 67.0%) compared to weekday admissions (33.6%), although this was not statistically significant at the 95% confidence level (p-value 0.06).
Table 34. The specialty of the first responsible consultant after onset of stroke

<table>
<thead>
<tr>
<th>Specialty</th>
<th>Weekday, % (admissions)</th>
<th>Weekend, % (admissions)</th>
<th>Total, % (admissions)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Neurology</td>
<td>2.89 (2 034)</td>
<td>2.70 (629)</td>
<td>2.84 (2 663)</td>
</tr>
<tr>
<td>Neurosurgery</td>
<td>4.31 (3 029)</td>
<td>4.42 (1 029)</td>
<td>4.33 (4 058)</td>
</tr>
<tr>
<td>Geriatric medicine</td>
<td>26.42 (18 583)</td>
<td>25.85 (6 022)</td>
<td>26.28 (24 605)</td>
</tr>
<tr>
<td>Accident and Emergency</td>
<td>3.06 (2 151)</td>
<td>3.36 (782)</td>
<td>3.13 (2 933)</td>
</tr>
<tr>
<td>General medicine</td>
<td>47.99 (33 749)</td>
<td>48.41 (11 277)</td>
<td>48.09 (45 026)</td>
</tr>
<tr>
<td>Other</td>
<td>15.33 (10 778)</td>
<td>15.27 (3 558)</td>
<td>15.31 (14 336)</td>
</tr>
</tbody>
</table>

7.5. Discussion

7.5.1.1. Summary of findings

We examined more than 90 000 patients admitted for stroke in England across a one-year period. The study demonstrated that patients admitted for stroke at the weekend have an increased risk of dying in hospital within the first 7 days. Moreover, the study also suggests that weekend care was different in terms of both access to treatments – with lower rates of urgent brain scans and thrombolysis – and wider outcome measures. Even after adjusting for various measures of case mix, weekend stroke admissions had higher rates of in-hospital complications (aspiration pneumonia) and a lower proportion of patients being discharged to their usual place of residence within 56 days.

7.5.1.2. Impact: magnitude of weekend effect

I calculated, as described in the methods section, that there are some 350 potentially avoidable in-hospital deaths within 7 days and an additional 650 people could be discharged to their usual place of residence within 56 days if the performance seen at weekdays was replicated at weekends.

There are also reasons to suggest that the inequality of care is more pronounced than identified in this study. If the effect is indeed caused by a staff deficiency and a lack of resources, then you would expect poorer quality and safety at all out-of-hours periods during the week, including bank holidays and weekday evenings and nights. A Dutch study, for instance, found stroke mortality rates were higher, in comparison with the Monday day shift, during Sunday and Monday evening shifts and during all night shifts.\(^{354}\) If this is the case, then the out-of-hours periods during weekdays are
masking some of effect. Our analysis was limited to out-of-hours days, since there is not a time stamp within the data.

7.5.1.3. **Amenable: association with resource levels**

There are several possible explanations for these findings, including: fewer clinical staff working in hospitals on weekends, with those who do work often having less experience and lower familiarity with the patients,\textsuperscript{253,244} and less accessibility of resources,\textsuperscript{351} such as radiologists to operate the scanners. This study provided some evidence to support the theory that one of the contributing factors to the ‘weekend effect’ might be a decreased proportion of stroke patients being admitted under the responsibility of stroke specialists at weekends. However, as ‘stroke’ is not a category of consultant specialty under the coding framework, the study had to define stroke-specialists as geriatricians, neurologists and neurosurgeons, as opposed to general medics, A&E consultants and other specialties. The association between consultant specialty and day of admission also provides some evidence that the ‘weekend effect’ is amenable to the provision of healthcare. To some extent, this finding is support by previous suggesting that stroke patients cared for by neurologists have better outcomes than those treated by other specialties.\textsuperscript{217}

7.5.1.1. **Amenable: experience of addressing the weekend effect**

More work is required to identify the cause of this ‘weekend effect’. However, looking at international experience provides evidence for optimism. Stroke care in Sweden in 1990-95 had similar levels of disparity to the current situation England with adjusted weekend/weekday odds ratios of 7-day in-hospital mortality of 1.14 (95% Confidence Interval 1.09 – 1.20) yet they managed to decrease the variation to 1.07 (1.01 – 1.14).\textsuperscript{358} The study also identified increases over this period in the proportion of patients discharged to their usual place of residence within 56 days.\textsuperscript{358}

7.5.1.1. **Amenable: reorganising stroke services**

Furthermore, solutions have been proposed to address the ‘weekend effect’. For instance, maintaining a more consistent level of activity – even if this requires hospital staff to be paid at higher levels during weekends – is sometimes economical.\textsuperscript{351} Previous studies have also demonstrated that it is possible to provide equal access to thrombolysis at weekends and that the use of comprehensive stroke centres and participation in stroke clinical improvement programmes can ameliorate the problem.\textsuperscript{356,359,360} In particular, regional reconfigurations represent a promising strategy for providing consistent levels of access to stroke care.
7.5.1.2. **Precision**

The indicators identified a wide range of performance, by day of admission, with statistically significant differences in five of the six indicators.

7.5.1.3. **Minimum bias**

HES data do not include information on the time of admission and therefore I was not able to investigate the wider issue of the quality of out-of-hours care in this study. Other limitations should also be noted. Firstly, the administrative database used gives only limited information on the severity of stroke, which has been suggested to be conceivably higher at weekends.\(^{362,363}\) Whilst I used a number of variables, such as Charlson index and number of previous admissions, to mitigate for any bias in case mix between admissions on different days of the week, there is no information on, for example, patients’ level of consciousness. A further limitation is that the administrative data can contain errors; however, you would not expect this to be different in the coding of weekday or weekend admissions.

I stratified one of the key results – in-hospital mortality – by age group to further investigate potential bias. Chances of death were higher for each of the six age groups for weekend admissions, which again provides some assurance that the results are not an artefact of bias. However, as outlined at the start of the results section, there are differing numbers of stroke patients on each day of the week which could suggest differences in average case-mix; if differences in severity are not recorded in the data then there is a risk of bias.

7.5.1.4. **Construct validity**

Five of the six indicators identified evidence of the weekend effect, and further analyses revealed lower level of specialist staffing at weekends, which provides evidence of construct validity. However, in lieu of having multiple comparable units – as with the cross-sectional analysis in the previous chapter – I have provided some assurance on validity by comparing consistency with previous literature.

The results are consistent with previous findings on the limited availability of stroke services at weekends in England, and that this is indeed reflected in lower levels of access. An audit of hospital care in 2010 highlighted that nursing practice was presumably leaving many patients in bed over a weekend which may be highly undesirable, including increasing the risk of venous thrombosis (blood clots). Furthermore, the same audit found that 43% of sites that provided thrombolysis services, either on or off site, were not able to provide the service 24/7.\(^{343}\) Similarly, a national study reported that despite improvements in access to urgent brain scanning, rates at weekends and evenings
remains significantly more limited. A study by Rudd and colleagues also found that patients were less likely to be admitted directly to a stroke unit (OR 0.77; 95% Confidence Interval 0.69–0.86) at weekends. Further to these findings on urgent stroke treatments, a recent report found that pathways allowing fast track transfer out of hospital covered only 59% of areas at weekends, compared to 75% at other times.

The findings are also consistent with some studies from other countries, including the surprising difference in the number of stroke admissions falling from Monday to Sunday, which in fact tallies with the results from a population based register of stroke incidence in Finland. A Swedish report also highlighted a lower proportion of patients admitted at weekends being discharged to their usual place of residence within 56 days (OR 0.96; 0.93 - 0.99). Similarly the rate of ischaemic stroke patients in a US study receiving screening for dysphagia (which could prevent aspiration pneumonia) falling from 6.1% on weekdays to 5.5% at weekends. The ‘weekend effect’ for 7-day in-hospital mortality (adjusted OR 1.18; 1.12 – 1.24) is also not dissimilar to studies from other countries, such as Sweden (1.07; 1.01 to 1.14), USA (1.14; 1.05 – 1.25), and Canada (1.14; 1.02 – 1.26).

7.5.1. Conclusion

Further work is needed to understand what organisational factors might influence the ‘weekend effect’ and to investigate centres that have reduced the disparities in access and outcome in out-of-hours care. A starting point for this is to allow hospitals to compare the extent of the ‘weekend effect’ in their organisation to that in their peers, as exemplified in Figure 17.
Figure 17. Hospital-level standardised mortality ratio, for deaths within 7 days of admission on a weekend in 2009/10

Note: Each dot represents a hospital. The horizontal line refers to national average; short-gauge dotted line refers to p<0.025 significance level; long-gauge dotted line refers to p<0.001 significance level.

There is also scope to extend this analysis to other specialties, with similar results having also been found in a limited number of other clinical areas, such as pulmonary embolism, hip fractures and upper gastro-intestinal haemorrhage.\textsuperscript{354} A greater understanding of the issue will also require better data, and the inclusion of an out-of-hours admission flag for hospital administrative data should be considered.

The analysis by day of week addresses most of the coding and risk-adjustment issues raised in the previous chapters and, therefore, permits more robust conclusions in what is, coincidentally, a very high impact area.
Part Three:

Obstetrics care
Chapter 8. Obstetrics literature review

<table>
<thead>
<tr>
<th>Current measurement</th>
<th>Case for specialty-specific use</th>
<th>Review of current indicators</th>
<th>Stroke</th>
<th>Application: validating common uses</th>
<th>Application: temporal analyses</th>
<th>Discussion</th>
</tr>
</thead>
<tbody>
<tr>
<td>Obstetrics</td>
<td></td>
<td></td>
<td>Stroke</td>
<td>Obstetrics</td>
<td>Obstetrics</td>
<td>Obstetrics</td>
</tr>
</tbody>
</table>

Discussion
PART III: Obstetrics Care

Overview

Context
The previous three chapters have shown that there is the potential to improve the application of stroke care indicators and, in doing so, to make meaningful inferences about quality and safety. This chapter – the first of three to focus on acute obstetric care – looks at how administrative data have been used to evaluate perinatal care – a very distinct area of hospital care to stroke treatment – and sets out how the indicators have been applied and validated to date.

Methods
A systematic review of original studies that applied, evaluated or validated obstetric-specific indicators based on administrative data. Indicator and study details were collated.

Findings
The initial search identified 1,670 unique citations of which 80 studies had used administrative data to evaluate perinatal care. There has been a steady growth in the use of such indicators; however, few studies have described the strategy used for identifying perinatal episodes from the data and there is a lack of consistency in the use of risk-adjustment of validation techniques. Many countries have used obstetrics-specific indicators although each study has only focused on a very small subset of measures. There is potential to make more meaningful inferences about the quality and safety of acute obstetric care by applying a more comprehensive set of indicators and being more explicit about the assumptions used.

What this chapter adds
The first review on obstetrics care focusing on: the use of administrative data; how these indicators have been used; and, importantly, how they have been applied and validated.
8.1. Background

8.1.1. Background on obstetrics

8.1.1.1. About obstetrics

Obstetrics, from obstare the Latin word for “to stand by”, is the specialty dealing with care of all women and their children during pregnancy, child birth and postnatal period. The delivery of obstetric care varies considerably from one nation to another; however, Figure 18 outlines key stages. Obstetrics care is delivered by a range of different healthcare professionals from country-to-country and may include hospital physicians and surgeons, community/family doctors, nurses, healthcare assistants or midwives.

Figure 18. Overview of the obstetric care pathway

8.1.2. Importance

8.1.2.1. Impact: global burden

Globally, there has been a general trend in improvement in obstetric outcomes with an average annual decline of maternal mortality (3.1%) and neonatal mortality (1.6%) in the 20 years from 1990. However, annual mortality rates remain at colossal levels, with at least: 3 million stillborn babies, 4 million neonatal deaths and half a million maternal deaths. Moreover, it has been suggested that the majority of these deaths are avoidable. Whilst poor outcomes are more prevalent in developing countries, it is far from limited to these nations. A study from the USA, published in 1992, found that 8% of the mothers or infants had serious complications in pregnancy, with twice this number being transferred for a higher-level of care either during or after delivery.

8.1.2.2. Impact: increasing risk

In these developed countries, in particular, there have been changes in the burden of obstetrics care with increasing numbers and complexity of deliveries. For instance, in England and Wales, overall fertility rates (which account for changes in over population numbers) increased by 21.5% in the decade to 2011 and by around 50% for women aged over 30.
8.1.2.3. Amenable to healthcare

There is good evidence to suggest that obstetrics outcomes can be improved through delivering better healthcare. For instance, one study found that problems of quality of care are the main cause of obstetric readmissions, while another study showed that continuous improvement efforts have shown to be effective for obstetric care in the UK.

8.1.3. Obstetrics measurement

8.1.3.1. Demand for indicators

A Cochrane review on maternal mortality and morbidity, which did not identify any RCTs on critical incident audit in any of the forty years preceding its publication, concluded that there is an essential need for feedback, directed to the relevant person. In England, a national audit report recently concluded that the Department has only limited assurance on the performance of maternity services due to a lack of comprehensive data on key outcomes and activity. The report recommended that a framework for gaining assurance should be developed.

8.1.3.2. Existing measures

The measurement of maternity outcomes is not new. For instance, in the UK, the current Confidential Enquiry into Maternal and Child Health can be traced to 1952, just four years after the inception of the NHS. A number of different methods have been introduced and, in 2006, Simpson and colleagues wrote a review of current methods for measuring perinatal patient safety. While flawed – for instance, there was no indication of how these current methods were identified – the paper does usefully set out the range of measurement techniques used to date, such as: safety attitude and climate surveys, focus groups, storytelling, executive walk rounds, and external review.

Indeed there are many existing quality measures of obstetrics. A recent review of guidance from the RCOG found 290 quality indicators covering 96 clinical categories, with a range of different definitions for each category. This exercise yielded a large number than a rapid literature review of the indicators on processes and outcomes of maternity services in Europe, Australia, New Zealand and the USA which identified 194 different indicators from 30 sources.

8.1.3.3. Using administrative data

In their review, Simpson et al suggest that the use of inexpensive, readily available administrative data to identify shortcomings in safety may provide a useful tool for identifying issues and monitoring trends in perinatal care and, went further by saying, “an on-going quality improvement program based on analysis of data from selected PSIs is essential”.

Yet, in the UK, there seems to
be little use of such indicators; of the 50 indicators relating to childbirth, none appear to use routinely collected hospital data as source. Results have been presented on a range using, for instance, radar plots and CUSUM charts to monitor obstetrician’s performance. 

8.2. Objectives

To produce a comprehensive list of indicators of the quality and safety of perinatal care that can be based on administrative hospital data, and provide a critical assessment of how these have been applied previously. This latter objective is designed to inform the quantitative analyses presented in the following two chapters.

8.3. Methods

8.3.1. Search resources and validation

An outline of the systematic methodology was given in a previous chapter (paragraphs 3.4.2 - 3.4.3, p72). The method used to search for peer-reviewed journal articles and other research is described earlier (paragraph 3.4.4, p74). An initial search of the databases – designed to identify peer-reviewed journal articles – was performed in June and July 2012, with a further search in February 2013.

8.3.2. Definitions

8.3.2.1. Definition for administrative data

Recall from paragraph 2.1.1.1 that ‘administrative data’ are defined, for the purpose of this review and to meet the ambition of the study, using the inclusion criteria: routinely collected data covering more than one hospital; used for administrative purposes (e.g. for reimbursement); and can be replicated using the internationally-recognised ICD9 or ICD10 diagnoses coding framework.

As a principle, where a study was based on administrative data linked to another supplementary source, the individual measures described in the paper were reviewed to assess whether they could be replicated with administrative data alone. This ensured that no potentially beneficial indicator was excluded and that validation studies where results from administrative data are compared to an augmented or different dataset were included.

5 Searched using the string: neonatal OR birth OR matern*
Studies based on registers of births and/or deaths without linkage to hospital episode data (such as diagnoses and procedures) were excluded.

8.3.2.1. Definition of quality and safety

As discussed earlier (paragraph 1.1.4), providing an exact definition of quality and safety is also difficult and its application for the purpose of this review on obstetric care is again best defined by the treatment of contentious indicators:

- Length of stay is included. There is an argument that length of stay is a poor measure of quality and safety: one study used length of stay instead as a measure of efficiency \(^{379}\), another found that it is not be statistically correlated with satisfaction \(^{380}\) and Oberer and Auckerman found that it is not inversely correlated with readmission rates. \(^{381}\) However, some variations – such as a dichotomous indicator of long length of stay – are more likely to be considered a proxy for aspects of quality.

- Emergency caesarean section rates are included, although they could be considered either a process measure (as caesarean section is an intervention initiated by a clinician) or an outcome of the care received earlier in labour (for example, emergency caesarean section following induced labour). \(^{375}\)

- Hospital charges is not considered as a quality indicator since this is a compound measure of patient characteristics, length of stay, and treatments provided – which can be picked up as separate indicators and risk adjustments – and is probably more obviously interpreted as a measure of hospital efficiency.

- Clinically associated complications of childbirth that are not amenable to healthcare – such as congenital malformations – are not included.

- Elective caesarean section (CS) rates are excluded as this is not related to a decision or intervention during the delivery admission. \(^{382}\)

8.3.2.2. Limitation to delivery admission

During the research it became evident that clinical care for the antenatal and postnatal aspects of the pathway have few agreed indicators of quality that can be applied to administrative data – due to limitations in the coding frameworks and relative lack of clinical guidelines – and clinical care provided is poorly recorded within English HES data. As a result, the obstetrics chapters focus on the delivery admission (or equivalent for home births) within the perinatal (intra-partum) stage.
Restricting the research to this aspect of the pathway does not affect the ability to meet the objectives of the project and, as well as being pragmatic, is grounded in evidence on:

- the influence of this aspect of the pathway on outcomes – an Italian review of neonatal deaths found that the great majority of deaths (87%) were due to perinatal causes; and

- activity and outcomes in perinatal care are less likely to be influenced by culturally-driven differences in decision making, which are likely to be more significant in ante- and post-natal care such as access to antenatal care.

### 8.3.3. Search strategy

#### 8.3.3.1. Electronic databases

The literature review was primarily based on searching the following databases:

- Ovid MEDLINE(R) In-Process & Other Non-Indexed Citations and Ovid MEDLINE(R) 1946 to Present

- Ovid Embase Classic+Embase 1947 to 2013 February 15

- Web of Science (to present)

#### 8.3.4. Search terms

The search strategy was informed by previous literature reviews relating to obstetric indicators. The first stage was to identify text words, synonyms and index terms for the studies on obstetrics care that applied, evaluated or validated quality and safety indicators based on routinely-collected administrative data. The synonyms were identified using the existing literature and MESH entry terms, a process suggested by the Cochrane review organisation (Table 35).

<table>
<thead>
<tr>
<th>Table 35. Search terms</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Key text words</strong></td>
</tr>
<tr>
<td>Medline index terms (MESH)</td>
</tr>
<tr>
<td>Embase index terms (EMTREE)</td>
</tr>
<tr>
<td>Birth/</td>
</tr>
<tr>
<td>Delivery/</td>
</tr>
</tbody>
</table>
PART III: Obstetrics Care

<table>
<thead>
<tr>
<th>Key text words</th>
<th>Obstetrics</th>
<th>Quality Indicators</th>
</tr>
</thead>
<tbody>
<tr>
<td>Synonyms for the text words (from MESH entry terms)</td>
<td>Parturition*</td>
<td>Quality indicator*</td>
</tr>
<tr>
<td>Birth*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Childbirth*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Obstetric*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Deliver*</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Other synonyms

Labour

Quality: quality; safe*; performance; medical error*; iatrogenic*; nosocomial*; adverse event*; sentinel event*; harm*

Indicators: administrative data; routine data; routinely collected data; claims data; international classification of diseases; international statistical classification of diseases; ICD*

Other spelling

Labor

International Classification of Diseases

The databases were interrogated using a two-dimensional search string using the ‘AND’ conjunction to identify papers relating to obstetric care which mention either a quality indicator or routine data. Medline, Embase and Web of Science were interrogated using the search string outlined in Figure 19 which was customised to the specific requirements of the search engines, searching the title, keywords, and abstract. Embase and Medline were also searched using indexing terms (EMTREE and MESH). For completeness non-English articles were included in the review.
8.3.4.1. Duplicates
Duplicates citations were identified using Endnote reference management software by identifying entries with identical authors, date, journal, and title.

8.3.5. Review process

8.3.5.1. First stage: Inclusion and exclusion criteria for review of indicator use
Recall that literature on quality and safety measures were collated if both: (1) based on routinely collected hospital administrative data; and (2) applied, evaluated or validated a potential or actual indicator of the quality or safety of hospital obstetric care. To meet this latter criterion, an included study had to provide numerical results on performance against the indicator.

Specifically, as a first stage, abstracts were reviewed and candidate papers were excluded if:

× None of the described indicators related to an perinatal care;

× None of the described indicators explicitly relate to the quality and safety of health care provided, such as measures of cost. As a result, readmission following birth is included but neither number of birth episodes nor those complications of births not associated with quality are included;
The study was not based on routinely-collected hospital administrative data (as defined above); and

The assessment of face validity of indicators which did not involve applying the algorithms to actual data did not meet the inclusion criteria.\textsuperscript{388}

Where there was insufficient detail in the abstracts to ascertain whether the criteria above were met, the study was included in the second stage (full text) review.

\textbf{8.3.5.2. First stage: Inclusion of additional papers describing indicators}

I also separately collated studies which described indicators that could be applied to administrative data but no actual database was used (and so failing the first stage criteria to provide numerical results) to ensure a well-defined comprehensive list of indicators. A key example of this is Dr Foster Intelligence’s consultation on obstetric measures (blood transfusion post-delivery rate; elective, emergency and total caesarean section rate; instrumental deliveries rate; maternal length of stay of 2 or 4 nights following normal or instrumental delivery) which described the indicators but did not give results on their application.\textsuperscript{389}

\textbf{8.3.5.3. Second stage review criteria}

For the second stage, the remaining articles had the same exclusion criteria applied to the full texts. This stage provided a form of quality assurance by double-reviewing many of the studies whilst also giving the opportunity to collate a range of additional information from the papers. As well as the details of the publication (such as author, title, year of publication) a standard template was used to extract additional information (Table 36). Information on validation techniques applied and any key limitations were also recorded where available. Information about the key results was also extracted. For the limited number of foreign language studies, these were translated by skilled colleagues.

\begin{table}[h]
\centering
\begin{tabular}{ll}
\hline
\textbf{Study section} & \textbf{Data field} \\
\hline
Design & Country \\
& Details of cohort \\
& Longitudinal, \textit{including months of data} \\
Data & Database(s) \\
& Year(s) of data \\
\hline
\end{tabular}
\caption{Information extracted}
\end{table}
8.3.5.4. Data extraction and review: analytical tools

I catalogued and reviewed the extracted abstracts using Endnote version X6 (EndNote Carlsbad, CA) reference management software. Statistical analyses of study details, such as trends in number of publications over time, were performed using Microsoft Office Excel 2010 (Microsoft Corp., USA).

8.3.5.5. Data extraction and review: Comparative analyses

In discussing the descriptive statistics on the papers identified, I use comparisons to previous literature reviews on patient safety indicators, predominantly that of Tsang et al\textsuperscript{66,67}, to provide a benchmark to judge whether the specialty-specific nature of the review, and the specialty in question (obstetrics), have revealed atypical results. A comparison to the stroke literature review (Chapter 4) is included in the discussion (Chapter 11).

The information from the studies was compared as a whole and in subsets since certain purposes, for instance, would require specific risk-adjustment models and strategies for identifying obstetric episodes. For example, where the purpose of the study is to validate the identification of safety issues against medical notes, it is unlikely that you would want to adjustment for case-mix where, for comparisons between providers, such a risk-adjustment would be important.

8.4. Results

8.4.1. Identification of studies that applied administrative data

8.4.1.1. Studies on application of obstetrics QIs

The initial search identified 1,670 articles of which 395 duplicates were removed. Of the remaining 1,275 articles screened, the full text was retrieved for 176 studies with the most common reason for exclusion being: non-relevant (n=283); the study describes an indicator using a non-administrative dataset (n=472); and evaluated only number of births (n=236). During the second review stage, a further 33 potential abstracts were added from the search of grey literature and review of the
bibliographies of key articles. Following the review of full texts, 129 citations were excluded resulting in 80 studies (Figure 20). A summary of the key characteristics is included in Figure 21. The details of a subset of papers are given in Table 37 on page 180.

**Figure 20. Inclusion and exclusion of articles in the literature review**

1,670 records identified through database searching

- Duplicate record?
  - Yes: 395 records excluded
  - No: 1,275 abstracts screened

- Eligibility criteria met?
  - No: 1,099 records excluded
  - Yes: 176 full-text articles assessed for eligibility; bibliographies reviewed

- References meet criteria?
  - Yes: 33 additional records identified through other sources and bibliographies of key articles
  - No: Eligibility criteria met?
    - Yes: 80 studies included for extraction of indicator details
    - No: 129 full-text articles excluded
### Figure 21. Characteristics of literature

<table>
<thead>
<tr>
<th>Publication year</th>
<th>Number of studies</th>
<th>Proportion of all studies</th>
<th>Example study</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pre-1995</td>
<td>0</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>1995-1999</td>
<td>1</td>
<td>1.3%</td>
<td>Kotagal (1999)</td>
</tr>
<tr>
<td>2005-2009</td>
<td>31</td>
<td>38.8%</td>
<td>Romano (2005)</td>
</tr>
<tr>
<td>Post-2009</td>
<td>30</td>
<td>37.5%</td>
<td>Srinivas (2010)</td>
</tr>
<tr>
<td>Not stated / not applicable</td>
<td>1</td>
<td>1.3%</td>
<td>BirthChoiceUK</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Country of study</th>
<th>Number of studies</th>
<th>Proportion of all studies</th>
<th>Example study</th>
</tr>
</thead>
<tbody>
<tr>
<td>USA</td>
<td>40</td>
<td>50.0%</td>
<td>Simonson (2007)</td>
</tr>
<tr>
<td>Australia</td>
<td>9</td>
<td>11.3%</td>
<td>Baghurst (2012)</td>
</tr>
<tr>
<td>England</td>
<td>8</td>
<td>10.0%</td>
<td>Bragg (2010)</td>
</tr>
<tr>
<td>France</td>
<td>4</td>
<td>5.0%</td>
<td>Bahrami (2008)</td>
</tr>
<tr>
<td>Taiwan</td>
<td>3</td>
<td>3.8%</td>
<td>Tseng (2010)</td>
</tr>
<tr>
<td>Other nations with n&lt;3 studies</td>
<td>15</td>
<td>18.8%</td>
<td>e.g. Canada in Thompson (2003)</td>
</tr>
<tr>
<td>Multi-national</td>
<td>1</td>
<td>1.3%</td>
<td>Drosler (2009)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Trend analysis</th>
<th>Number of studies</th>
<th>Proportion of all studies</th>
<th>Example study</th>
</tr>
</thead>
<tbody>
<tr>
<td>Longitudinal</td>
<td>19</td>
<td>23.8%</td>
<td>Linton (2004)</td>
</tr>
<tr>
<td>Non-longitudinal</td>
<td>60</td>
<td>75.0%</td>
<td>Fantani (2006)</td>
</tr>
<tr>
<td>Detail not available/not applicable</td>
<td>1</td>
<td>1.3%</td>
<td>BirthChoiceUK</td>
</tr>
</tbody>
</table>

| Months of data, mean (IQR) | 40.7 (12-68) | - | |

<table>
<thead>
<tr>
<th>Diagnosis coding</th>
<th>Number of studies</th>
<th>Proportion of all studies</th>
<th>Example study</th>
</tr>
</thead>
<tbody>
<tr>
<td>ICD-9</td>
<td>40</td>
<td>50.0%</td>
<td>Romano (2005)</td>
</tr>
<tr>
<td>ICD-10</td>
<td>16</td>
<td>20.0%</td>
<td>Bottle (2009)</td>
</tr>
<tr>
<td>ICD-9 and -10</td>
<td>2</td>
<td>2.5%</td>
<td>Lutomski (2012)</td>
</tr>
<tr>
<td>Not described</td>
<td>18</td>
<td>22.5%</td>
<td>Keino (2012)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Number of indicators</th>
<th>Number of studies</th>
<th>Proportion of all studies</th>
<th>Example study</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>38</td>
<td>47.5%</td>
<td>Baghurst (2012)</td>
</tr>
<tr>
<td>&gt;1</td>
<td>42</td>
<td>52.5%</td>
<td>Reilly (2004)</td>
</tr>
</tbody>
</table>

**TOTAL** 80 -

### 8.4.2. Details from studies applying administrative data

#### 8.4.2.1. Country

Half of studies (n=40; 50.0%) were from the USA, with only four other nations (Australia, England, France and Taiwan) identified as having 3 or more studies included. There was one international study, which covered seven countries, although a further study compared the results from the authors’ country (England) to that of previously published results the USA. The USA predominantly
PART II: Obstetrics Care

use ICD-9 diagnosis coding frameworks and, as a result, the majority of studies included used this coding framework, within only 16 explicitly using the ICD-10 framework.

8.4.2.2. Year of study
I did not identify any studies applying administrative data meeting the criteria that were published from 1998 or earlier, even though the databases were searched from their inception to present. There has been a strong approximately linear trend (Figure 22, $\beta = 0.7$) in annual publications, rising from one in 1999 to thirteen in 2012.

Figure 22. Trend in annual publication numbers

8.4.2.3. Study types
The majority of studies were cross-sectional in design, with a limited number of other designs, such as: case-control studies (n = 4) to, for instance, understand the burden of safety incidents, and quasi-experimental studies (n=4) designed to evaluate the effect of the implementation of a policy. One of these studies showed that the reporting of performance data may have a positive impact on quality improvement. The papers also included studies that could be described as having cohort designs, with hospital type or day of admission as the exposure.

8.4.2.4. Purpose
The studies covered a wide range of uses. Some compared performance between different delivery units, whether: physician; hospitals; regions; and countries. Other studies, however, used sub-group analysis to evaluate differences in performance, with key variables of:
- hospital characteristics – including location, size, teaching status and ownership \(^{83,406,415,425-429}\) – and including performance against other measures, such as blood supply at the facility, \(^{430}\) level of obstetrical specialist support, \(^{431}\) anaesthetist staffing level, \(^{396}\) and caesarean section rate \(^{432}\);

- patient characteristics \(^{83,396,399,400,402,433-437}\) including type of insurance \(^{391}\);

- delivery risks, such as gestational age \(^{438}\), type of delivery \(^{405,439}\), and length of stay \(^{390,398}\);

- temporal aspects, such as year-on-year changes \(^{411,440-443}\), evaluation of the “July” phenomena, \(^{444}\) and day of admission \(^{445}\);

Nine studies looked at the overall incident rate \(^{81,370,446-449}\) and their burden \(^{407-409}\), with sixteen focusing on validation. \(^{392,393,410,439,449-460}\) Four studies evaluated change in performance from pre- to post- implementation of a new policy, \(^{412-414}\) including the impact of IT. \(^{461}\)

### 8.4.3. Details from datasets

#### 8.4.3.1. Data source

The studies reviewed used a range of different administrative databases, with the most frequent single dataset (n= 9) being the Nationwide Inpatient Sample in USA, which includes all patients from a stratified random sample of non-federal hospitals in 19-28 states. The State Inpatient Databases (SID), which are also from the AHRQ-sponsored Health Cost and Utilization Project (HCUP), were also used widely.
### Table 37. Example papers

<table>
<thead>
<tr>
<th>Article author (year)</th>
<th>Country, data year</th>
<th>Study design</th>
<th>Sample</th>
<th>Measures</th>
<th>Results</th>
<th>Covariates</th>
<th>Key subject</th>
<th>Related articles</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zhan et al (2003)</td>
<td>USA, 2000</td>
<td>Case-control</td>
<td>720,021 (neonates)</td>
<td>Birth trauma</td>
<td>0.7%</td>
<td>Age, sex, case-mix (DRG), SES, insurance, comorbidities, hospital characteristics</td>
<td>Burden of incidence</td>
<td>AHRQ PSIs</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Obstetric trauma (VBwI)</td>
<td>22.4%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Obstetric trauma (VBwol)</td>
<td>8.6%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Obstetric trauma (CS)</td>
<td>0.7%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cameron et al (2007)</td>
<td>Australia, 1993-2002</td>
<td>Cross-sectional</td>
<td>775,073</td>
<td>PPH</td>
<td>5.6%</td>
<td>Age, mode of delivery</td>
<td>Year</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>PPH readmissions</td>
<td>1.1%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Neonatal deaths</td>
<td>0.3%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Readmissions within 10 days</td>
<td>10.4%</td>
<td>Day of admission, Prenatal care visits</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Emergency department visits within 21 days</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bottle et al (2009)</td>
<td>England (2005-06)</td>
<td>Cross-sectional</td>
<td>15,298</td>
<td>Obstetric trauma (VBwI)</td>
<td>60.5%</td>
<td>Age</td>
<td>Hospital comparison</td>
<td>AHRQ PSIs</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Obstetric trauma (VBwol)</td>
<td>27.9%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Obstetric trauma (CS)</td>
<td>2.9%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lutomski et al (2011)</td>
<td>Ireland</td>
<td>Cross-sectional</td>
<td>649,019</td>
<td>PPH</td>
<td>2.6%</td>
<td>Age, marital status, medical card status, SES, comorbidities, delivery complications, mode of delivery</td>
<td>Delivery type; Year</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Blood transfusion rate</td>
<td>1.3%</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
8.4.3.2. Case ascertainment

Only a minority of studies (26.1%) described the strategy for identifying birth/delivery episodes from the administrative data. Of these, a range of methods were used, such as: specific ICD diagnosis codes\(^ {455} \), diagnosis-related group (DRG)\(^ {455} \), procedure codes\(^ {397,411} \) or a mixture of these strategies.\(^ {454,414,424,446} \) Across just those studies which used the less commonly applied ICD-10 data, the following identifiers were used: procedure codes (OPCS) R17-R25; DRG and diagnosis codes; OPCS R17 (elective CS only); and presence of ICD-10 Z37 (outcome of delivery) flag.

8.4.3.1. Extract validation

A small number of studies evaluated or explicitly discussed coding accuracy, including: purposefully identifying coding errors;\(^ {434} \) accounting for variations in how comprehensively the data is entered (coding depth);\(^ {401,455} \) and making more general descriptive discussion on changes in coding practice.\(^ {442} \) One study explicitly evaluated the influence of the strategy for identifying birth episodes within the administrative data identification strategy.\(^ {454} \)

8.4.4. Risk models

8.4.4.1. Case mix adjustment

When evaluating whether an algorithm based on administrative data is accurate in identifying actual events one might expect that non-risk-adjusted rates would be used but, even when excluding these validation studies, there were still many which did not use risk adjustment.\(^ {394,401,407,413,420,422,430,434,442,446} \)

Where risk-adjustment was used, the maternal factors that were included as covariates ranged from a single factor (age)\(^ {404} \) to applying numerous maternal factors (such as age, ethnicity, proxy for socio-economic status, parity, comorbidities and prior CS\(^ {397} \)). Other studies applied infant risk factors (e.g. birth-weight\(^ {390,391,396,403,405,417,418,426,431,438,439,445,460} \) and congenital malformation\(^ {393,403} \), delivery risks (e.g. mode of delivery\(^ {390,391,405,433,440,443,446} \) and complications during delivery\(^ {395,403,425,436,454,461} \) and other proxies for complexity (e.g. day of admission\(^ {433} \) or hospital characteristics\(^ {391,395,399,401,409,425,436,441,443,444,450,454,461} \)).

Even for the same indicators, there was little consistency in the use of risk-adjustment models. For instance:

- CS rates – ranged from no adjustment\(^ {422} \) to age, citizenship, insurance, comorbidities, pregnancy risks, birth-weight, congenital malformations, delivery complications\(^ {403} \)
PART III: Obstetrics Care

- Obstetric trauma rates – ranged from no adjustment\(^{401}\) or just age\(^{404}\) to a combination of age, case-mix (DRG), proxy for socio-economic status, insurer, and comorbidities\(^{409}\)

8.4.4.2. Participant exclusions

Again there was little consensus in the exclusion of patients from the denominator, even when the same potential bias was being mitigated against. For instance, in trying to exclude high-risk deliveries, some studies used gestational age as a cut off but, even in these instances, they ranged in the cut off mark (from 24 weeks to 37 weeks).\(^{411,415,438}\) Other approaches included setting a minimum birth-weight\(^{420}\), excluding deliveries after transfer\(^{429}\), or excluding non-cephalic deliveries\(^{375}\). Some studies also excluded deliveries to hospitals with fewer than a certain threshold of annual cases.\(^{425,430}\) Many studies excluded certain modes of delivery, such as vaginal or caesarean, with some of these limitations reflecting the scope of the study rather than purposeful efforts to reduce bias.

8.4.5. Details from indicators

8.4.5.1. Range of indicators

The literature review also identified studies, that whilst not applying actual data so not meeting the inclusion criteria of the primary review, described an indicators which is proposed for application to administrative data. As set out in the methods, these studies were retained separately to ensure that the resulting list of indicators was comprehensive and well described. Such studies included the papers recommending the use of AHRQ indicators\(^{40}\), SimPatIE\(^{104}\) and OECD HCQIs\(^{85}\).

A large number of studies were based on AHRQ’s indicators – consisting either just AHRQ indicators (using either all or a subset of obstetrics measures)\(^{81,83,401,404,407-409,416,437,439,441,451,455}\) or using explicitly AHRQ indicators alongside other measures\(^{393,436}\). A full list of indicators is included at Appendix E, p306 with some key measures included in Table 38.

Table 38. Key obstetrics indicators

<table>
<thead>
<tr>
<th>Pathway</th>
<th>Outcome measures</th>
<th>Example study</th>
</tr>
</thead>
<tbody>
<tr>
<td>Maternal</td>
<td>Emergency readmissions</td>
<td>Thompson (2003)(^{400})</td>
</tr>
<tr>
<td></td>
<td>Puerperal sepsis</td>
<td>Lu (2005)(^{427})</td>
</tr>
<tr>
<td></td>
<td>Obstetric trauma</td>
<td>Downey (2012)(^{441})</td>
</tr>
<tr>
<td></td>
<td>Episiotomy</td>
<td>Webb (2002)(^{418})</td>
</tr>
<tr>
<td></td>
<td>Mortality</td>
<td>Magalhaes (2012)(^{446})</td>
</tr>
</tbody>
</table>
8.4.5.2. Indicator validation

The majority of studies (n=42) studies for which full-text was available did not evaluate the validity of the indicators applied. Where used, the most common techniques were:

1. construct validation, which is based on the hypothesis that measures of similar aspects should be statistically, positively associated; and

2. validation of the risk adjustment model (minimum bias).

The construct validation – which also inherently assesses data accuracy – took the form of comparison to either a “gold standard”, such as case notes, or another dataset. A similar method involved comparing results derived from the same administrative dataset by applying different algorithms for the same adverse outcome (e.g. trauma). Forms of validation of the risk adjustment model were similarly diverse and included simply reporting model fit performance, comparison of different model specifications, or evaluating reliability of the model by applying additional data. For one study, the main focus of the research was to develop a valid model for predicting an adverse outcome.

Less common strategies for validation were to assess reliability by evaluating performance for different cohorts (e.g. including or excluding breech presentations in the denominator) and to follow a multi-faceted validation process (e.g. following the procedure set out by AHRQ for their indicators). In one study, the hospital-level results were sent to the organisations for qualitative feedback, with hospital trusts “generally positive”.

<table>
<thead>
<tr>
<th>Post-partum haemorrhage</th>
<th>Goff (2012)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Severe maternal morbidity (composite)</td>
<td>Callaghan (2012)</td>
</tr>
<tr>
<td><strong>Delivery</strong></td>
<td></td>
</tr>
<tr>
<td>Unsuccessful vaginal birth after caesarean (VBAC)</td>
<td>Knight (2013)</td>
</tr>
<tr>
<td>Emergency caesarean section</td>
<td>Roberts (2009)</td>
</tr>
<tr>
<td>Failed vacuum or forceps delivery</td>
<td>Knight (2013)</td>
</tr>
<tr>
<td>Late caesarean section</td>
<td>Knight (2013)</td>
</tr>
<tr>
<td><strong>Neonate</strong></td>
<td></td>
</tr>
<tr>
<td>Infections</td>
<td>Korst (2005)</td>
</tr>
<tr>
<td>Birth trauma – injury to neonate</td>
<td>Downey (2012)</td>
</tr>
<tr>
<td>Adverse outcome indicators</td>
<td>Walker (2010)</td>
</tr>
</tbody>
</table>
8.4.5.3. Indicator interpretation

“An example of a commonly used maternity indicator that is difficult to interpret is the overall caesarean section rate. Lower caesarean section rates are often assumed to reflect better care. However, there is also a threshold below which the caesarean section rate is too low and babies may be harmed. One problem is that there are no established guidelines for determining this threshold. A second problem is that as elective caesarean sections become increasingly popular,” (p1) 375 which may increase rates – irrespective of the intervention of the health services – where there is a policy of maternal choice.

8.5. Discussion

8.5.1. Principle findings

8.5.1.1. Descriptive data on studies

I identified 80 studies that described the application, evaluation or validation of indicators of obstetrics care using routinely collected hospital data. While there were no studies meeting the inclusion criteria published prior to 1999, there has been an approximately linear increase since then. The majority of studies were based on data from the USA and only a third (n=21, from 61 where diagnosis framework stated) used the ICD-10 coding framework. Approximately a quarter of studies used the longitudinal nature of the administrative data to make inferences about quality and safety of care. Around half of studies (38 of 80) used only one indicator. The studies took many forms, from descriptive cross-sectional studies of differences in performance for various hospital and patient characteristics to quasi-experimental designs to evaluate the effect of the implementation of an initiative to improve care.

Some studies have used composite indicators, such as the Maternity Unit Performance Index that include both mother and child outcomes as used by Anastasakis.463 There are some potential advantages in bringing together different facets of quality into a concise quantified measure; however, within the literature I found little explanation over whether these indicators are instead heavily influenced by the sub-measure with the largest range in performance.

8.5.1.2. Case ascertainment strategies for births

The review of studies identified a lack of consistency – both in terms of which fields, and codes within the chosen fields, are used – in the identification of birth episodes within the administrative data. These studies are also not necessarily consistent with other purposes with, for instance:
- a study led by York University on obstetrics efficiency - using length of stay as a proxy – that used a combination (ICD10: any code beginning "Z37"; OPCS4.5: any code beginning "R17", "R18", "R19", "R20", "R21", "R22", "R23", "R24", "R25"), and
- a recent publication on coding accuracy used admission classification: admimeth 82 (other: babies born in healthcare provider) and 83 (other: babies born outside the healthcare providers, except when born at home as intended). This could represent an issue if the episodes that are not consistent to two different strategies are not wholly representative of the shared cases within the cohort. Taking the latter example, which used admission method as a flag, the strategy excludes ‘other birth events’ which captures NHS-funded home births and all other birth events which are not NHS funded, either directly or under an NHS service agreement. Previous research has shown that results for patient characteristics and outcomes are not necessarily the same for home births and, therefore, differing treatment (in terms of inclusion/exclusion in study population) of this group would potentially introduce bias. Some excluded transfers as proxy for high-risk; however increases in midwifery-led care, which have high transfer rates, may invalidate this assumption.

8.5.2. Indicator validity

8.5.2.1. Risk-adjustment

Achieving good outcomes for the increasingly complex caseload is problematic and sometimes will not be amenable to the quality of healthcare provided. For instance, in the UK between 1985 and 1993, 55% of maternal deaths occurred despite being judged to have received high quality medical care. Given this evolving complexity in obstetrics care, adjusting performance measures for case-mix is increasingly important. Yet the literature review revealed a lack of transparency and consistency. Disaggregating the results further highlights this problem. In the 2007 indicator set described by the SimPatIE project, there were four obstetrics indicators, with no stratification of results, and yet different depth of risk adjustment, from none to accounting for age and comorbidities (Table 39). No rationale for the differences in sophistication between the different risk models was presented.
Table 39. Risk adjustment in the SimPatIE obstetrics measures

<table>
<thead>
<tr>
<th>Indicator</th>
<th>Risk adjustment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Obstetric trauma - Vaginal Delivery without instrument</td>
<td>Age and comorbidities</td>
</tr>
<tr>
<td>Obstetric trauma - Vaginal Delivery with instrument</td>
<td>Age</td>
</tr>
<tr>
<td>Obstetric trauma – Caesarean section</td>
<td>None</td>
</tr>
<tr>
<td>Problems with Child Birth</td>
<td>None</td>
</tr>
</tbody>
</table>


Other studies have mentioned the importance and current issues with risk models but then fail to address these. For instance, Janakiraman and Ecker\textsuperscript{467} mention that risk adjustment of birth trauma is not well established; however, the authors do not take the next logical step to recommend that this should be empirically tested – as is possible using the administrative data. Indeed such analysis has been done elsewhere with one study – although using a neonatal specific data source – evaluating the risk adjustment explicitly using a systematic review of literature to identify potential risk factors.\textsuperscript{468} Whilst having a good predictive risk model is important for the purpose of evaluating quality and safety, it is also has wider importance such as to ensure that activity-based payment schemes are fair.\textsuperscript{469}

8.5.2.2. Indicator validation

Whilst some indicators have been accepted for over a decade – such as the AHRQ PSIs, including obstetrics measures, that were accepted in 2002\textsuperscript{82} – the validation is embryonic. As discussed earlier there is also a lack of consistency in the techniques and comprehensiveness in the validation of the obstetrics indicators presented here. The European-wide SimPatIE project only applied an assessment of face validity, involving reaching a consensus amongst a small group of clinicians, before recommending their implementation.\textsuperscript{388} Despite the relative ease with which it can be done using administrative data, few studies evaluated construct validity (by comparing results from similar measures). Indeed, one study evaluated the AHRQ PSIs by comparing performance (using Spearman’s rank) across different indicators but did not include the obstetrics measures.\textsuperscript{470}

There are also examples of recommendations being made with only limited knowledge of the validity of the indicators. In recommending indicators for future use, the SimPatIE project repeated a conclusion that “the AHRQ PSIs are a broad screen for potential safety events that point to needed improvement in the quality of care for specific population”; however, this was based on a study that looked only at patient characteristics rather than, say, hospital or temporal covariates.\textsuperscript{104}
While this reliance on using clinical consensus over definitions goes some way to meeting two of the three aspects in Freidman’s model on what makes a good indicator – proxy power (meaningful to health) and communication power (transparent) – it fails to address data power (ensuring high quality and available data). Moreover, as discussed earlier, in reality the application of the data often lacks the required transparency to be judged as having reasonable communication power.

8.5.2.3. Performance

The heterogeneity of the studies – for instance in their use of risk-adjustment and strategy to identifying birth episodes – means that formal meta-analyses of results were not possible. As well as synthesising results and comparing against peers it is also possible to compare indicators performance to a predetermined benchmark and, whilst not an a priori aim of this part of the study, reviewing the literature revealed that some acceptable standards for these indicators have been stated (Table 40). This adds to the evidence that perinatal care is amenable (as certain performance can/should be achieved).

Table 40. Selected benchmarks for performance

<table>
<thead>
<tr>
<th>Measure</th>
<th>Acceptable range</th>
<th>Cohort</th>
<th>Organisation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Caesarean section rates</td>
<td>5 – 15%</td>
<td>All births in the population</td>
<td>UNICEF 1997\textsuperscript{472}</td>
</tr>
<tr>
<td>Post-Partum Haemorrhage</td>
<td>≤4%</td>
<td>Numerator: post-partum haemorrhage ≥1000ml within 2 hours post-partum</td>
<td>Danish National Indicator Project \textsuperscript{473}</td>
</tr>
<tr>
<td>Maternal mortality</td>
<td>≤1%</td>
<td>Proportion of women with obstetric complications admitted to a facility who die</td>
<td>UNICEF 1997\textsuperscript{472}</td>
</tr>
<tr>
<td>Third or fourth degree lacerations</td>
<td>≤6%</td>
<td>All first time vaginal deliveries</td>
<td>Danish National Indicator Project \textsuperscript{473}</td>
</tr>
</tbody>
</table>

8.5.3. Strengths and limitations

8.5.3.1. Strengths

Previous studies – including one in 2013 – had brought together clinically recommended indicators from medical associations, such as the Nordic obstetric and gynaecological association (NOGA) and
American College for Obstetricians and Gynaecologists (ACOG), RCOG (UK), Royal Australian & New Zealand College Obstetrics and Gynaecology (RANZCOG), Society for Obstetricians and Gynaecologists of Canada (SOGC). However, this current review represents the most comprehensive assessment of how indicators have been applied using administrative data and adds important transparency to the issues over their applications, such as on risk-adjustment models. The comprehensive nature from this study stems from its scope and also the process followed such as the inclusion of extracts from all languages.

As with the stroke review, the specialty-specific literature review process was found to be a sensitive strategy for identifying relevant research and, as such, provides more assurance that the resultant citations are comprehensive. For instance, a literature review of all patient safety measure based on routinely collected hospital data, which followed a similar process, identified a similar magnitude of studies (n = 124) despite its far broader scope.

8.5.3.2. Limitations of the review process

No formal meta-analysis could be undertaken because of the heterogeneity in study methodologies and underlying data frameworks. I did not use a template to assess the quality of the papers since there was a very diverse range of purposes across the studies and, as already discussed, validity needs to be proportionate to the purpose. Given the strong evidence on the existence of publication bias, studies using indicators which did not identify significant results when assessing the quality and safety of perinatal care may have been missed by the search strategy in this study. However, the design of the study – to retain articles which described indicator sets, but did not apply those to administrative data – mitigates against this bias to some extent.

The inclusion criteria that the indicators should measure performance of perinatal care excluded studies that looked at general obstetric performance, such as measures of obstetric nosocomial infections and obstetric readmissions. However, these indicators were applied to perinatal care elsewhere so are included in the list of potential indicators.

8.5.3.3. Limitations of results

Maternity care differs from other areas of health care in a number of ways – such as by predominantly dealing with healthy people – and, therefore, the findings might not be directly generalisable to other specialties. Also maternity services deal with the culturally and emotionally sensitive area of childbirth and so non-biomedical outcomes, which are not captured in the administrative datasets used in the studies presented here, would also need to be considered to have a full understanding of the overall performance of obstetric care.
The results do not give a comprehensive set of indicators of obstetric care nor results on this. The scope of the review, which was necessitated by a number of factors described above (paragraph 8.3.2.2, p170), also meant that the measures were focused on the perinatal aspect of the obstetrics care pathway and, as such, the indicators described could not give a holistic picture on the overall performance of obstetric services. For instance, neonatal indicators (which in many health services will be the responsibility of a separate specialty: neonatology) including nosocomial infections from neonatal units and ambulatory care sensitive conditions which are inherently important aspects of quality and safety are not included. A specific reason for the potential incompleteness of the list of indicators is that some sets of measures are not based on routinely collected data such as the commonly applied trigger tools based on chart reviews.

8.5.4. Conclusions

8.5.4.1. Contribution to knowledge of indicators

As with the literature review of stroke indicators (Chapter 4) looking at a specialty in detail, had advantages over a broader scope; not least, to better identify existing indicators. The study has highlighted a paucity of use of such indicators and a lack of transparency and consensus on key assumptions. The importance of the transparency added by this review was reiterated by previous calls for consistent definition and reporting practices for hospital complications across countries.

8.5.4.2. Purpose

A set of quality and safety measures would be valuable in improving obstetrics care. However, such an indicator set should be meaningful, actionable, timely, feasible and presented data to caregivers, which are all plausible requirements within the constraints of routinely-collected hospital data. Recommendations on exactly how these indicators should be used vary. Many applications have been suggested for use by organisations such as AHRQ, JACHO, OECD, Australian Council for Safety and Quality in Health Care (ACSQ) and Canadian Institute for Health Information (CIHI); however, these recommendations have not necessarily been echoed by the academic community who have conducted much of the research to date. For instance, Kristensen and colleagues recommended the use of obstetric trauma-vaginal delivery without instrument (AHRQ, JCAHO), with instrument (AHRQ, CIHI, OECD) and birth trauma-injury to neonate (AHRQ, CIHI, OECD) for use, whereas the same study suggested that obstetric trauma following caesarean delivery (AHRQ, OECD) and problems with childbirth (ACSQ, OECD) should not be used.
Whilst administrative data have been used in published literature to identify nosocomial infections, there is not universal agreement that this is an appropriately sensitive surveillance tool.\textsuperscript{457} Table 41 sets out key recommendations for future obstetrics research based on the findings of the review.

### Table 41. Summary of implications of study design

<table>
<thead>
<tr>
<th>Finding</th>
<th>Implication</th>
<th>Recommendation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Obstetrics indicators based on administrative data have been used infrequently outside the USA and with ICD-10 coding framework</td>
<td>Barriers to generalising findings across borders mean that little is known about the use of such indicators.</td>
<td>Key study designs should be replicated using data from outside USA or Canada to identify any potential outliers.</td>
</tr>
<tr>
<td>2. There is a lack of consensus and transparency about the assumptions used in applying the indicators to administrative data</td>
<td>Barrier to generalising, validating or benchmarking results.</td>
<td>To explicitly set out the assumptions used in studies, such as the choice of diagnosis codes and fields.</td>
</tr>
<tr>
<td>3. Around half of studies used only one indicator</td>
<td>Content validity of many existing studies questionable</td>
<td>Apply bundle of indicators</td>
</tr>
<tr>
<td>4. Only two studies evaluated the effect of coding practice or accuracy on performance and these focused solely on coding depth</td>
<td>Concerns remain over the validity of results given the potential influence of coding bias.</td>
<td>To analyse the potential effect of coding variation.</td>
</tr>
<tr>
<td>5. Only one study – USA-based with limited scope – evaluated the effect of day of admission on performance.</td>
<td>Potentially important quality/safety issue is being ignored.</td>
<td>A non-USA study on day of admission should be conducted to identify whether there is a potential issue.</td>
</tr>
</tbody>
</table>
Chapter 9. Obstetrics variation and validation
Overview

Context
The previous chapter reviewed the current use of administrative data to evaluate the quality and safety of obstetric care in hospitals. This chapter sets out how a subset of the indicators identified can be applied, using data from hospitals in England. As well as evaluating whether outliers can be identified (precision), and to address the gap in the literature identified in the previous chapter, the following study will also measure the effect of differences in coding and investigate other aspects of validity.

Methods
Eight indicators, spanning process and outcomes measures for both the mother and neonate, were applied to English data in 2011/12. Logistic regression analyses were used to adjust for case-mix, and funnel plots to look at potential outliers. I also investigated the effect of variation in hospitals’ coding practice using a range of proxies and multi-level regression models.

Findings
The indicators were sufficiently sensitive to identify variations in performance as recorded in the data, with 441 occurrences of hospital trusts – or 550 outliers at the site-level – performing statistically differently than expected at the 99.8% significant level across the eight indicators. Differences in coding practice appeared to only partially explain the variation in performance. The work showed that administrative data can be used to indicate potential concerns around performance across some of the different aspects of quality and safety of obstetric care.

What this chapter adds

This chapter represents the most comprehensive review of obstetric care in England using administrative data. This study is also the first to evaluate the effect of variation in hospital- and site-level coding of such measures.

Acknowledgements
Professor Susan Bewley (King’s College, London) commented on potential indicators.
9.1. Background

9.1.1. Obstetrics care in England

9.1.1.1. Policy development in maternity care

In England, having a baby is the most common reason for admission to hospital, with some 700,000 births currently each year. In total, maternity care costs around £2.6 billion annually which accounts for nearly 3% of the NHS budget. Maternity services are relatively unique in that they are focused on supporting predominantly healthy people through a natural life event.

Maternity care has changed dramatically over time. In the 1950s, around a third of births took place at home; however, following a trend of increased hospitalisation during the 1960s and 1970s, this proportion currently stands at around 2% nationally. There have been improvements in outcomes with, for example, the number of babies dying within 28 days of births falling from 6.7 to 2.9 per 1,000 live births in the 30 years from 1981. However, there is evidence to suggest that outcomes in obstetrics in England can be improved further, including:

- that some outcomes are consistently worse than in the other UK nations and wider international comparisons; and
- the failure to fully address the common causes for litigation claims, such as errors during caesarean section, with the cost to cover against maternity claims totalling for around a fifth of the total cost of maternity services.

9.1.1.2. Existing measures of obstetrics care

The policy for the commissioning and delivery of maternity services was defined by the Department of Health in Maternity Matters (2007), which recommended the development of a range of indicators, that will enable users, trusts and commissioners to compare performance against others. However, as recently as 2013, a publication from the Royal College of Obstetricians and Gynaecologists highlighted that “there has been an astonishing lack of robust information on even simple clinical outcomes on a national basis.” (p.viii)

9.1.2. Revisiting existing literature

9.1.2.1. Previous hospital-level comparisons

To date, there have been few studies that have used administrative data to compare obstetric performance at a hospital level, with the existing studies focusing on caesarean section or obstetric trauma rates (Figure 23). These studies do cover a range of countries’ health services,
including developed and developing countries. However, there have also been comparative studies focusing on a different delivery unit, whether physicians, regions, or countries. Other studies have also used sub-group analysis to focus on specific hospital characteristics, such as:

- location, size, teaching status and ownership, including performance against other measures, such as bloody supply at the facility;
- level of obstetrical specialist support;
- anaesthetist staffing level; and
- caesarean section rate.

A recent report by the Royal College of Obstetricians and Gynaecologists introduced a number of indicators of maternity indicators for English providers. However, the work did not look at neonatal outcomes – rather focusing solely on data from the maternal patient record – and did not empirically test the validity of the measures.

9.1.2.2. Data and indicator validation revisited

As discussed in the previous chapter, there have been a small number of studies investigating coding accuracy. For instance, on identifying coding errors by reviewing medical records and coding depth, with Drosler et al using the mean number of secondary codes (although not presenting the data on the obstetrics measures) as a proxy for coding practice, and Grobman et al using quartile of ICD-9 coding intensity (although this is metric is poorly defined in the paper). Other studies on data validation include a more descriptive discussion on changes in coding practice and an investigation into the strategy for identifying birth episodes within the administrative data identification strategy.

Few studies have explicitly validated the indicator algorithms, with the techniques used being primarily comparison to other measures of performance and validation of the risk adjustment model. The former primarily took the form of criterion validation; namely comparison to a validated measure, such as medical records. Forms of validation of the risk adjustment model included simply reporting model fit performance, comparison of different model specifications, or evaluating reliability of the model by applying to additional data. For one study, the main focus of the research was to develop a valid model for predicting an adverse outcome.
Figure 23. Studies comparing hospital performance using obstetric QIs

<table>
<thead>
<tr>
<th>Cohort</th>
<th>Country</th>
<th>Measures</th>
<th>Risk-adjustment</th>
<th>Main results</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Caesarean section rates</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Turner (2011) 420</td>
<td>Irish</td>
<td>CS rates (elective and emergency)</td>
<td>None</td>
<td>Overall CS rates ranged from 18.7 – 35.6%</td>
</tr>
<tr>
<td>Bragg et al (2010) 397</td>
<td>England</td>
<td>CS rate (elective and emergency)</td>
<td>Age, ethnicity, SES, and clinical risk factors for CS</td>
<td>Adjusted CS rates ranged from 13.6 – 31.9%</td>
</tr>
<tr>
<td>Fantini et al (2006) 403</td>
<td>Italy</td>
<td>CS rates (primary)</td>
<td>24 risk factors in the full model; 7 (marital status, maternal age, infant weight, and clinical risk factors) in the parsimonious model.</td>
<td>Correlation between hospital rankings of two models was 0.92</td>
</tr>
<tr>
<td>Hsu (2006) 421</td>
<td>Taiwan</td>
<td>CS rates</td>
<td>Maternal and paternal age and education, gestational age, multiple births, parity, infant gender, clinical risk factors</td>
<td>33.7% rate. Substantial change in rankings for crude vs risk adjusted models.</td>
</tr>
<tr>
<td>Gichangi (2001) 422</td>
<td>Kenya</td>
<td>CS rates</td>
<td>None</td>
<td>Hospital-based rates of 6.3% (range 0.3 – 37.7%)</td>
</tr>
<tr>
<td><strong>Obstetric trauma</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Baghurst (2013) 419</td>
<td>Australia</td>
<td>Third and fourth degree lacerations</td>
<td>Maternal age, parity, ethnicity, assistance with instruments and episiotomy, shoulder dystocia, infant birth-weight</td>
<td>8 hospitals with systematically more/fewer tears</td>
</tr>
<tr>
<td>Bottle et al (2009) 404</td>
<td>England</td>
<td>Third and fourth degree lacerations</td>
<td>Age</td>
<td>60.5% (with instrument), 29.7% (without instrument), 2.9% (CS). Part of wider study using AHRQ PSIs.</td>
</tr>
<tr>
<td>Webb et al (2002) 418</td>
<td>USA</td>
<td>Episiotomy rates</td>
<td>“Case mix” factors including maternal age, race/ethnicity, education and infant birth-weight</td>
<td>Hospital episiotomy rates ranged from 20 to 73%, and lacerations from 4 to 13%. Rates of two indicators correlated (p&lt;0.01)</td>
</tr>
</tbody>
</table>
9.1.2.3. **Revisiting limitations in existing literature**

The review of literature described in the previous chapter suggested that future research should focus on:

1. applying key study designs for identifying any potential outliers outside the USA or Canada;
2. using a more comprehensive set of indicators;
3. being explicit about the assumptions used in studies, such as the choice of codes and fields for identifying delivery episodes;
4. analysing the potential effect of coding variation; and
5. investigating the association between quality and day/month of delivery outside the USA to identify whether there is a potential issue.

The latter area is covered in the next chapter while the other areas are covered in this section.

9.2. **Objectives**

In this chapter I investigate the feasibility of using HES data to evaluate the quality and safety of obstetric care at a provider level, focusing on the ability of measures to identify important effects in a robust fashion.

9.3. **Methods**

9.3.1. **Data**

There are some noteworthy differences in the data collected for maternity care in comparison to the other areas of healthcare contained in the HES dataset, which are outlined here and discussed in further detail at the end of the chapter. HES contains data for all births in England, including those at home and in non-NHS hospitals, irrespective of whether a consultant or midwifery-led episode. Pregnant women are admitted as general admissions and only once they have given birth does their record become a maternity record.

There are two types of maternity record – (1) a delivery record for the mother, and (2) the birth record(s) for the baby (*or babies in the case of multiple deliveries*) – with each containing an additional 19 fields covering, for instance, parity of mother, method of onset of labour, place of delivery, gestational age and birth-weight. The data in these records comes from birth notification
records and requires only a limited dataset to be completed; however, these additional fields are not as complete as the rest of the HES data.\textsuperscript{488,489} Despite including both maternal and baby records, the HES dataset do not link the baby’s record to that of the mothers, or vice versa.

9.3.2. Selection of indicators

\textbf{9.3.2.1. Selection criteria revisited}

Recall the criteria listed earlier (paragraph 3.5.1.2, p76) for the selection of indicators. For this obstetric element to the project, the experts involved in getting a consensus on the indicators are listed in paragraph 9.3.3.1 below.

\textbf{9.3.2.2. Selection of indicators}

The literature review described in the previous chapter identified 24 maternal, 9 delivery, 11 neonatal categories of indicators with the potential to be applied to administrative data in England. Following the selection criteria, a subset of eight indicators was chosen for application here (Figure 24). The indicators were chosen to cover both process (e.g. early caesarean section rates) and outcome (e.g. infection rates). These two examples also demonstrate that the chosen indicators span care given during the delivery admission, from decision on type of delivery through to postpartum care.
9.3.3. Developing indicator definitions

9.3.3.1. Development process

As with the stroke measures, the selected indicators were then refined using a multi-faceted approach. In the first instance, the indicator definitions were implied from any previous studies that applied these measures:

1. I derived draft definitions using details from algorithms applied in existing literature, identified through the systematic review outlined in the previous chapter.

2. I used an iterative consensual approach to ascertain feedback from a multidisciplinary panel of experts, consisting of a senior obstetrician, a public health doctor, and administrative data experts (from Imperial’s Dr Foster Unit).

3. The definitions were also reviewed by a multi-disciplinary team, including people experienced in the management of health services and clinicians, as part of the review of the process of publishing a national report on maternity services.¹⁵²

9.3.3.2. Final definitions

The finalised indicator definitions are included below in Table 42.

Table 42. Details of obstetric indicators used in analysis

---

Note: the positioning of the complications is for indicative purposes only.
<table>
<thead>
<tr>
<th>Indicator</th>
<th>Rationale</th>
<th>Definition (discharges meeting the inclusion/exclusion criteria for the denominator with...)</th>
<th>Exclusions</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Delivery</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>a Early (pre-39 week) caesarean section</td>
<td>NICE recommends¹ that planned caesarean section should not routinely be carried out before 39 weeks due to increased risk of respiratory morbidity.</td>
<td>Elective caesarean sections (OPCS code: R17) where gestational age &lt; 39 weeks</td>
<td>Diagnoses possibly justifying early caesarean section (ICD-10 codes: O100-9, O11, O140-9, O16, O244/9, O266/8, O356, O360/1/5, O410/1, O421, O430, O440, O694, O991/4/8, Z352)²</td>
</tr>
<tr>
<td><strong>Maternal</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>b Perineal tear</td>
<td>Third and fourth degree tears during vaginal delivery are not all preventable, but the risk can be reduced through appropriate labour management and care standards.</td>
<td>Admission includes both diagnosis (ICD-10: O702-3) and procedure code (OPCS: R322, R325)</td>
<td>Caesarean sections</td>
</tr>
<tr>
<td>c Puerperal infection</td>
<td>Puerperal sepsis is associated with poor care (especially hygiene) at the time of birth</td>
<td>Admission includes diagnosis code for puerperal sepsis (ICD-10: O85-6) within 42 days of birth</td>
<td>None</td>
</tr>
<tr>
<td>d 30-day emergency readmission (all cause)</td>
<td>Readmission rates have previously been shown to be associated with the quality of hospital care</td>
<td>Emergency inpatient admission within 30 days, where ‘admission source’ field does not suggest transfer from other provider</td>
<td>Births where the discharge date is in final month of year covered in data year (March)</td>
</tr>
<tr>
<td><strong>Neonatal</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>e In-hospital perinatal mortality</td>
<td>Still births plus in-hospital deaths within 7 days</td>
<td>Value suggesting still born in birth status (2, 3, 4) or discharge fields (5) fields, and discharged within 7 days with discharge method (4) suggesting death</td>
<td>None</td>
</tr>
<tr>
<td>f Injury to neonate</td>
<td>Although rare, birth trauma to the neonate is often preventable³</td>
<td>Injury diagnosis code within admission (ICD-10: P102-4, P108-12, P114-5, P119, P122, P130-1, P138-9, P142, P148-9, P15)</td>
<td>Premature births, as identified through diagnosis code (ICD-10: P070-3), gestational age (&lt;28 weeks) or birth-weight (&lt; 2.5kg). Injury of skeleton (Osteogenesis imperfecta diagnosis, ICD-10: Q780) Stillborn babies</td>
</tr>
<tr>
<td>g Selected neonatal infections</td>
<td>Neonatal infection is associated with poor care (especially hygiene) at the time of birth</td>
<td>Diagnosis procedure code (ICD-10: P36, P372/5/8-9, O753, O85-6, A41, A32, A49)</td>
<td>Premature births, as identified through diagnosis code (ICD-10: P070-3), gestational age (&lt;28 weeks) or birth-weight (&lt; 2.5kg). Stillborn babies</td>
</tr>
<tr>
<td>h 28-day emergency readmissions (all cause)</td>
<td>Readmission rates have previously been shown to be associated with the quality of hospital care</td>
<td>Emergency inpatient admission within 28 days, where ‘admission source’ field does not suggest transfer from other provider</td>
<td>Births where the discharge date is in final month of year covered in data year (March) Missing birth discharge date (c. 1% of births) Death during birth admission</td>
</tr>
</tbody>
</table>

Notes:

9.4. Application and validation of indicators

9.4.1. Assumptions

9.4.1.1. Case ascertainment considerations

The next stage was to extract both baby and maternal records from 1 April 2011 to 31 March 2012. There are many potential fields and codes within HES to identify records relating to the delivery admission (Figure 25).

To evaluate the importance of the case ascertainment algorithm, I checked the consistency of coding across the different fields. Some of these are governed by coding rules, with examples given in Figure 26 and whilst these rules were adhered to in 100% of cases in 2010-11 (used for exploratory purposes), across both birth and delivery records, I identified there were potential issues for using the ‘epitype’ field for case ascertainment.
### Figure 26. Coding rule examples

<table>
<thead>
<tr>
<th>Rule</th>
<th>Detail</th>
<th>Example issue</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rule 150</td>
<td>Where date of birth and start of episode are not missing, and date of birth equals epistart, then the first episode in the spell (epiorder = 1) with admission method 82 (Other: babies born in health care provider) is set as epitype = 3 (birth episode)</td>
<td>Does not apply for admimeth = 83 (Other: babies born outside the healthcare provider, except when born at home as intended) which were coded as: epitype = 3 (birth episode) for 57.7% (2709) of episodes; with 0.0% (2) as epitype = 2 (delivery episode) and 42.2% (1982) as epitype = 1 (general episode).</td>
</tr>
<tr>
<td>Rule 160</td>
<td>Where certain treatment codes (first 3 chars of oper_01 = R17 - R25) are recorded then episode set as epitype = 2 (delivery episode)</td>
<td>Does not cover secondary diagnoses (fields oper_02 to oper_20).</td>
</tr>
</tbody>
</table>

Source: HES Data Dictionary

#### 9.4.1.2. Case ascertainment algorithm

I decided that a more detailed algorithm would be required than using the ‘epitype’ as a flag for a delivery or birth episode. After comparing the potential sensitivity and specificity of the possible case ascertainment algorithms and reviewing similar algorithms used in the existing literature, such as Knight et al (2013) and Murray et al (2012), I proceeded with the codes outlined below. I did not use an algorithm based on procedure codes for identifying birth episodes since this strategy lacked sensitivity as the completeness of coding of these fields has been questionable, particularly historically.

**Delivery episodes:** Valid delivery method in either procedure (oper_01 to oper_20 = R17 to R25) OR maternity tail (delmeth = 0 to 9).

**Birth episode:** epitype = 3, 6 OR admimeth = 82, 83.

As with the stroke analysis, episodes were grouped into superspells – representing the continuous period that a patient receives care and includes transfers between hospitals. Where analyses were undertaken at the hospital/unit level, performance against the quality and safety measure was attributed to the first provider of care.
9.4.1.1. Unit of analysis: trust- and unit-level

To investigate the variation in performance across England, an appropriate level of analysis is required. In 2007, the Department of Health reiterated the policy that women should have a choice of where to give birth and, as a result, providers are expected to offer multiple different maternity units. For instance, many NHS trusts have both a hospital-based consultant-led obstetric unit and a separate midwifery-led unit on a different site. Taking this example of an organisational set up, analysis at the unit-level will be better able to investigate the effect of organisational and structural decisions at this unit-level, whereas it is more likely to be affected by bias due unmeasured differences in case-mix resulting from different in the admission criteria to different units (a freestanding midwifery unit is typically intended for low risk mothers only). As such, I compared the two provider fields (‘procode5’ and ‘sitetret’) to an existing list of providers from a national audit of maternity services. To validate the matching, I checked that the annual number of births recorded in the two datasets were similar. I report the analyses at both the trust and unit-levels.

9.4.2. Risk adjustment

9.4.2.1. Overview of case-mix adjustment

Once an extract had been obtained using the above criteria, the first step was to identify outliers, which involved applying the indicator definitions to obtain denominators and numerators for each hospital. A logistic regression was used to calculate an expected number of numerator events based on the case-mix for each trust/unit to account for case mix factors. In addition, some of the measures have specific case-mix factor considerations as outlined in Table 42. The methods used are described in more detail in paragraphs 3.6.2 to 3.6.6.

9.4.3. Validation

9.4.3.1. Statistical testing and testing construct validity

Crude and standardised rates were plotted using funnel plots with 95% and 99.8% control limits and identified outliers (see paragraph 3.6.8, p81). Hospitals' performance across the different indicators was compared – at both trust and unit-levels – to evaluate the hypothesis that certain indicators would be correlated. This involved 28 pairwise comparisons for each of the institutional levels, with further detail on this method included in the discussion below. Given the lack of process measures, I supported the evaluation of construct validity by also exploring the association between

---

1 Age of mother, sex of baby, parity (maternal indicators only), multiple deliveries, socio-economic deprivation (carstairs quintile), previous caesarean section (maternal only), ethnic group, gestational age, birthweight, delivery method and other maternal conditions (Pre-existing diabetes, gestational diabetes, pre-existing hypertension, pre-eclampsia or eclampsia, placenta praevia or abruption, polyhydramnios, oligohydramios.)
performance and intervention rates (using standardised instrumental rates). Since these were hypothesised as being correlated and not an un-hypothesised statistical trawling exercise, Bonferroni correction or other adjustment to p-values for multiple comparisons was not used, although I highlight where relationships are significant at the 99.8%, as well as 95%, levels.

9.4.3.2. Sensitivity analysis for coding practice

There is a risk that hospitals’ performance for these indicators might be largely affected by variation in the way they code their data rather than due to differences in quality and safety. As such, the final stage of analysis was to investigate, at a site level, the consistency of coding practice and evaluate the relationship between any coding bias and hospital performance. Coding completeness was investigated, by comparing to performance in a relevant indicator, through:

1. Including a measure of ‘coding depth’, with the coding practice in some sites increasing the likelihood that they record secondary diagnoses and, therefore, identify complications and comorbidities.

2. Excluding sites with poor coding practice (defined as those sites with fewer than 90% of cases with both birth-weight and gestational age fields completed), and evaluating whether a similar proportion are still identified as outliers.

The effect of the first of these variations in the modelling [1], involved evaluating the effect of hospital-level variable to account for coding completeness, and so I fitted multi-level generalised linear models and again plotting on a funnel chart to identify outliers.

9.4.4. Amenable: associations between organisational factors and performance

9.4.4.1. Amenable: size and performance

To test whether the quality of care being measured by the indicators is amenable to how care is delivered, I undertook some exploratory analysis to investigate whether organisational factors were associated with variations in performance. In the first instance, I compared unit size (number of births) with performance against the indicators. This should be interpreted with caution as tertiary or midwifery-led units are likely to be different sizes to general maternity units and will treat a different case-mix.
9.4.4.2. Amenable: midwifery staffing levels

I also used data requested from the Health and Social Care Information Centre (HSCIC) on midwifery staffing levels, by trust and month. The number of births and maternities each month were calculated and a trust-level indicator of relative staffing levels, ‘S’, was defined as:

\[ S_{i,t} = \frac{D_{i,t} - \sum D_{i,t}}{\sum M_{i,t}} \]

where \( i = \) trust \( i; \) \( t = \) month \( \epsilon \{ \text{April 2011, \ldots, March 2012} \}; \)

\( D = \) number of deliveries (births or maternities); \( M = \) number of midwives

I used this approach rather than comparing birth-to-midwife ratios between providers, since the latter is likely to reflect differences in type of provider (indeed staff-to-bed ratio have previously been used as a proxy to differentiate between major teaching, minor teaching, and nonteaching hospitals\(^{257}\)).

9.4.5. Alignment with analytical framework

The table below summarises how this chapter aligns with the overall analytical framework (Table 43).
Table 43. Analytical framework

<table>
<thead>
<tr>
<th>Domain</th>
<th>Study design</th>
<th>Method described</th>
<th>Results described</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Impact</td>
<td>Descriptive, national rates</td>
<td>-</td>
<td>9.5.1.1</td>
</tr>
<tr>
<td>2. Amenable</td>
<td>Descriptive, range</td>
<td>-</td>
<td>9.5.3</td>
</tr>
<tr>
<td></td>
<td>Staffing levels</td>
<td>9.4.4.2</td>
<td>9.5.2.1</td>
</tr>
<tr>
<td>3. Face and content validity</td>
<td>Consultation</td>
<td>9.3.3.1</td>
<td>-</td>
</tr>
<tr>
<td>4. Precision</td>
<td>Hospital results displayed on funnel plots</td>
<td>9.4.3.1</td>
<td>9.5.3</td>
</tr>
<tr>
<td>5. Construct validity</td>
<td>Correlations between the set of indicators, including outcome v. process indicators</td>
<td>9.4.3.1</td>
<td>9.5.4</td>
</tr>
<tr>
<td>6. Minimum bias</td>
<td>Risk-adjustment model</td>
<td>9.4.2</td>
<td>9.5.5.1</td>
</tr>
<tr>
<td></td>
<td>Sensitivity analyses of influence of coding error and practice.</td>
<td>9.4.3.2</td>
<td>9.5.5.2, 9.5.5.3</td>
</tr>
<tr>
<td>7. Data availability</td>
<td>Application of indicators to show ICD-10 based data is sufficient.</td>
<td>By implication</td>
<td></td>
</tr>
<tr>
<td>8. Reporting burden</td>
<td>Application of indicators.</td>
<td>By implication</td>
<td></td>
</tr>
<tr>
<td>9. External and ecological validity</td>
<td>Explicit statement of assumptions used.</td>
<td>9.4.1</td>
<td></td>
</tr>
</tbody>
</table>

9.4.6. Analytical packages

All regression analyses were conducted using SAS version 9.2 using either the PROC LOGISTIC or PROC GLIMMIX procedures. Funnel plots were created using the template provided by the Association of Public Health Observatories (APHO, [http://www.apho.org.uk](http://www.apho.org.uk)).

9.5. Results

9.5.1. Importance

9.5.1.1. Maternal characteristics

I identified 669,617 maternities and 678,785 live and still births in the period April 2011 to March 2012. Some 61.5% of maternities were delivered by spontaneous vertex, with the characteristics of the study population in Table 44. The table also includes details of the cohort when excluding sites with poor coding, as described in the methods section (paragraph 9.4.3.2, p203). In addition other maternal complications included: pre-existing hypertension (0.5% of women), pre-existing diabetes
(0.6%), gestational diabetes (3.2%), pre-eclampsia/eclampsia (1.9%), placenta praevia/abruption (1.0%), polyhydramnios (1.0%), oligohydramnios (0.9%).

Table 44. Characteristics of mothers in the study population

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Categories</th>
<th>Maternities, %</th>
<th>Restricted cohort</th>
<th>p-value (restricted vs excluded cohort)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>n = 669,616</td>
<td>n = 516,919</td>
<td></td>
</tr>
<tr>
<td><strong>Delivery method</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Spontaneous vertex</td>
<td></td>
<td>61.5</td>
<td>61.7</td>
<td></td>
</tr>
<tr>
<td>Spontaneous, other cephalic</td>
<td></td>
<td>0.5</td>
<td>0.5</td>
<td></td>
</tr>
<tr>
<td>Low forceps, non-breach</td>
<td></td>
<td>4.2</td>
<td>4.4</td>
<td></td>
</tr>
<tr>
<td>Other forceps, non-breach</td>
<td></td>
<td>2.2</td>
<td>2.0</td>
<td></td>
</tr>
<tr>
<td>Ventouse, vacuum extraction</td>
<td></td>
<td>6.3</td>
<td>6.2</td>
<td></td>
</tr>
<tr>
<td>Breech</td>
<td></td>
<td>0.4</td>
<td>0.5</td>
<td></td>
</tr>
<tr>
<td>Breech extraction not otherwise specified</td>
<td></td>
<td>0.1</td>
<td>0.1</td>
<td></td>
</tr>
<tr>
<td>Elective caesarean</td>
<td></td>
<td>10.1</td>
<td>9.9</td>
<td></td>
</tr>
<tr>
<td>Emergency caesarean</td>
<td></td>
<td>14.7</td>
<td>14.8</td>
<td></td>
</tr>
<tr>
<td>Other than those specified above</td>
<td></td>
<td>0.1</td>
<td>0.1</td>
<td></td>
</tr>
<tr>
<td><strong>Maternal age</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt;19</td>
<td></td>
<td>5.0</td>
<td>5.1</td>
<td></td>
</tr>
<tr>
<td>20-24</td>
<td></td>
<td>18.5</td>
<td>18.8</td>
<td></td>
</tr>
<tr>
<td>25-29</td>
<td></td>
<td>27.8</td>
<td>27.9</td>
<td></td>
</tr>
<tr>
<td>30-34</td>
<td></td>
<td>28.9</td>
<td>28.7</td>
<td></td>
</tr>
<tr>
<td>35-39</td>
<td></td>
<td>15.8</td>
<td>15.5</td>
<td></td>
</tr>
<tr>
<td>40 and over</td>
<td></td>
<td>4.0</td>
<td>4.0</td>
<td></td>
</tr>
<tr>
<td><strong>Ethnicity</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>White</td>
<td></td>
<td>72.1</td>
<td>72.4</td>
<td></td>
</tr>
<tr>
<td>Asian</td>
<td></td>
<td>10.5</td>
<td>10.5</td>
<td></td>
</tr>
<tr>
<td>Black (inc. Black British)Afro-Caribbean</td>
<td></td>
<td>5.0</td>
<td>5.2</td>
<td></td>
</tr>
<tr>
<td>Mixed</td>
<td></td>
<td>1.5</td>
<td>1.5</td>
<td></td>
</tr>
<tr>
<td>Other (inc. Chinese)</td>
<td></td>
<td>3.3</td>
<td>3.4</td>
<td></td>
</tr>
<tr>
<td>Unknown / not stated</td>
<td></td>
<td>7.6</td>
<td>7.0</td>
<td></td>
</tr>
<tr>
<td><strong>Level of deprivation</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td></td>
<td>14.9</td>
<td>14.2</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td>16.1</td>
<td>15.9</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td>18.9</td>
<td>18.8</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td>22.1</td>
<td>22.2</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Parity</td>
<td>p &lt; 0.0001</td>
<td></td>
<td></td>
</tr>
<tr>
<td>---------------</td>
<td>-------------</td>
<td>------------</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Primiparous</td>
<td>43.2</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Multiparous</td>
<td>56.8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Parity</td>
<td></td>
<td>p &lt; 0.0001</td>
<td></td>
<td></td>
</tr>
<tr>
<td>'6' (unknown)</td>
<td>1.3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1.4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gestational age</td>
<td>&lt; 37</td>
<td>7.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>8.6</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>37-39</td>
<td>34.9</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>38.7</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>40-41</td>
<td>42.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>46.7</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>&gt;=42</td>
<td>3.7</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>4.0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Unknown</td>
<td>11.9</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1.9</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Birth-weight</td>
<td>&lt;2500g</td>
<td>5.8</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>6.4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2500-4000g</td>
<td>74.3</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>80.1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>&gt;4000g</td>
<td>10.2</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>11.0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Unknown</td>
<td>9.7</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1.7</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Multip</td>
<td>Yes</td>
<td>1.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1.5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>No</td>
<td>91.7</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>97.1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Unknown</td>
<td>6.9</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1.4</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Notes: \( \chi^2 \) test for association, excluding 'unknown' categories

### 9.5.1. National level results

Of those women meeting the inclusion criteria for the maternal complications: 3.1% had a perineal tear, 0.9% had an infection, and 1.0% were readmitted within 30 days of discharge. Of those neonates meeting the inclusion criteria for the relevant complications, 0.7% were stillborn or died in-hospital within 7 days, 1.5% were injured, 2.1% had an infection, and 6.1% were readmitted within 28 days of discharge. Around a third (33.9%) of the 67,370 elective caesarean sections were recorded as being conducted prior to 39 weeks gestation; after excluding deliveries for which there was no recorded gestational age (15.4%) and/or a recorded indication for early caesarean section (20.7%) this equates to 35.2% of cases.

### 9.5.2. Amenable

#### 9.5.2.1. Association with organisational factors

The range in performance is shown in Table 46, p211. The association between staffing level and complication rates suggested that lower staffing levels were associated with higher perinatal mortality (\( p = 0.05 \)) while there were not significant correlations with the other measures (Table 45).
PART III: Obstetrics

Table 45: Association between birth-to-midwife ratio and performance measures

<table>
<thead>
<tr>
<th>Label</th>
<th>Indicator</th>
<th>Effect</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>B</td>
<td>Perineal tear</td>
<td>-0.109</td>
<td>0.336</td>
</tr>
<tr>
<td>C</td>
<td>Puerperal infection</td>
<td>0.083</td>
<td>0.603</td>
</tr>
<tr>
<td>E</td>
<td>In-hospital perinatal mortality</td>
<td>0.397</td>
<td>0.054</td>
</tr>
<tr>
<td>F</td>
<td>Injury to neonate</td>
<td>0.228</td>
<td>0.107</td>
</tr>
<tr>
<td>G</td>
<td>Selected neonatal infections</td>
<td>0.134</td>
<td>0.277</td>
</tr>
</tbody>
</table>

Notes: Readmission rates were not calculated as the performance could be affected by staffing levels in subsequent month. Early caesarean section was not calculated as this was not assumed to be affected by midwifery staffing levels.

9.5.3. Precision

Displaying the performance at the provider level on funnel plots, across either the 149 trusts or 229 units, highlighted the variation in performance (Figure 27 and Figure 28). All the indicators identified at least trusts/sites with performance outside the 99.8% control limits (Table 46). The number of births per trust ranged from 1,291 to 10,782, and for sites from 10 to 8,516.
Figure 27: Funnel plots of trust-level performance across the obstetric indicators, 2011-12

Note: Each dot represents a trust. The horizontal line refers to national average; short-gauge dotted line refers to p<0.025 significance level; long-gauge dotted line refers to p<0.001 significance level.
Figure 28: Funnel plots of site-level performance across the obstetric indicators, 2011-12

Note: Each dot represents a site. The horizontal line refers to national average; short-gauge dotted line refers to $p<0.025$ significance level; long-gauge dotted line refers to $p<0.001$ significance level.
### Table 46: Number of trusts identified as having statistically significant above or below average performance

<table>
<thead>
<tr>
<th>Label</th>
<th>Indicator</th>
<th>National rate</th>
<th>Variation</th>
<th>Outliers: number of providers with different performance to average (p&lt;0.001)$^1$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Range, by trust$^2$</td>
<td>Trust-level (n = 149)</td>
</tr>
<tr>
<td>a</td>
<td>Early caesarean section</td>
<td>35.2%</td>
<td>[1.1% – 60.7%]</td>
<td>12</td>
</tr>
<tr>
<td>b</td>
<td>Perineal tear</td>
<td>3.1%</td>
<td>[1.0% – 5.3%]</td>
<td>13</td>
</tr>
<tr>
<td>c</td>
<td>Puerperal infection</td>
<td>0.9%</td>
<td>[0.2% – 1.9%]</td>
<td>22</td>
</tr>
<tr>
<td>d</td>
<td>30-day maternal emergency readmissions</td>
<td>1.0%</td>
<td>[0.2% – 3.0%]</td>
<td>15</td>
</tr>
<tr>
<td>e</td>
<td>In-hospital perinatal mortality</td>
<td>0.7%</td>
<td>[0.3% – 1.0%]</td>
<td>5</td>
</tr>
<tr>
<td>f</td>
<td>Injury to neonate</td>
<td>1.5%</td>
<td>[0.1% – 6.9%]</td>
<td>37</td>
</tr>
<tr>
<td>g</td>
<td>Selected neonatal infections</td>
<td>2.1%</td>
<td>[0.2% – 11.1%]</td>
<td>34</td>
</tr>
<tr>
<td>h</td>
<td>28-day neonatal emergency readmissions</td>
<td>6.1%</td>
<td>[1.6% – 14.1%]</td>
<td>31</td>
</tr>
</tbody>
</table>

**Note**

1: Some trusts/sites were removed for specific indicators (the total number given represents the maximum)

2: Range represents minimum and maximum performance for providers, for unadjusted rates

3: Better care is indicated by lower than average performance for these measures.
PART III: Obstetrics

9.5.4. Construct validity

Across the trust-level indicator results, there were five pairs of indicators which had a statistically significant correlation at the 95% level, of which one was significant at the 99.8% level (Table 47).

### Table 47: Trust-level Coefficient of correlation between pairs of indicators

<table>
<thead>
<tr>
<th></th>
<th>Perineal tear</th>
<th>Puerperal infection</th>
<th>30-day maternal emergency readmissions</th>
<th>In-hospital maternal emergency readmissions</th>
<th>Injury to neonate</th>
<th>Selected neonatal infections</th>
<th>28-day neonatal emergency readmissions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Early caesarean section</td>
<td>-0.122</td>
<td>-0.076</td>
<td>-0.049</td>
<td>0.005</td>
<td>-0.189*</td>
<td>-0.077</td>
<td>-0.135</td>
</tr>
<tr>
<td>Perineal tear</td>
<td></td>
<td>0.121</td>
<td>0.198**</td>
<td>-0.119</td>
<td>-0.107</td>
<td>-0.174*</td>
<td>-0.227*</td>
</tr>
<tr>
<td>Puerperal infection</td>
<td></td>
<td></td>
<td>-0.067</td>
<td>0.014</td>
<td>0.218*</td>
<td>0.190*</td>
<td>-0.121</td>
</tr>
<tr>
<td>30-day maternal readmissions</td>
<td></td>
<td></td>
<td></td>
<td>0.002</td>
<td>0.080</td>
<td>-0.022</td>
<td>0.048</td>
</tr>
<tr>
<td>In-hospital perinatal mortality</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.038</td>
<td>0.008</td>
<td>-0.069</td>
</tr>
<tr>
<td>Injury to neonate</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.062</td>
<td>0.074</td>
</tr>
<tr>
<td>Selected neonatal infections</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>-0.012</td>
</tr>
</tbody>
</table>

Note 1: Correlation significant at the 95% level marked with ‘*’; those at 99.8% level with ‘**’

In further exploratory analysis, I found that higher standardised instrumental rates are associated with higher rates of: emergency caesarean section ($0.451, p < 0.0001$); maternal infection rates ($0.225, p = 0.007$); and tears ($0.38, p < 0.0001$). Across the site-level indicator results, there were seven pairs of indicators which had a statistically significant correlation at the 95% level, of which three were significant at the 99.8% level (Table 48).
Table 48: Site-level coefficient of correlation between pairs of indicators

<table>
<thead>
<tr>
<th></th>
<th>Perineal tear</th>
<th>Puerperal infection</th>
<th>30-day maternal emergency readmissions</th>
<th>In-hospital perinatal mortality</th>
<th>Injury to neonate</th>
<th>Selected neonatal infections</th>
<th>28-day neonatal emergency readmissions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Early caesarean section</td>
<td>-0.107</td>
<td>-0.014</td>
<td>-0.030</td>
<td>0.006</td>
<td>-0.076</td>
<td>0.030</td>
<td>-0.004</td>
</tr>
<tr>
<td>Perineal tear</td>
<td>0.036</td>
<td>0.012</td>
<td>0.081</td>
<td>0.153*</td>
<td>-0.045</td>
<td>-0.202*</td>
<td></td>
</tr>
<tr>
<td>Puerperal infection</td>
<td></td>
<td></td>
<td>0.129</td>
<td>0.067</td>
<td>0.259**</td>
<td>-0.051</td>
<td></td>
</tr>
<tr>
<td>30-day maternal</td>
<td></td>
<td></td>
<td></td>
<td>0.218**</td>
<td>-0.024</td>
<td>0.179*</td>
<td>0.074</td>
</tr>
<tr>
<td>readmissions</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>In-hospital perinatal</td>
<td></td>
<td></td>
<td></td>
<td>0.047</td>
<td>0.355**</td>
<td>0.153*</td>
<td></td>
</tr>
<tr>
<td>mortality</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Injury to neonate</td>
<td></td>
<td></td>
<td></td>
<td>0.017</td>
<td></td>
<td>-0.064</td>
<td></td>
</tr>
<tr>
<td>Selected neonatal</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.097</td>
<td></td>
</tr>
<tr>
<td>infections</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note 1: Correlation significant at the 95% level marked with ‘*’; those at 99.8% level with ‘**’

9.5.5. Minimum bias: coding practice

9.5.5.1. Effect of case-mix adjustment

The analysis of the association between adjusted and unadjusted rates, suggested that the case-mix adjustment had the least effect on the measure of maternal readmissions (Pearson correlation coefficient, \( r = 0.969 \)) and greatest effect on perinatal mortality (\( r = 0.540 \)) (Table 49). These results tally with the model fit statistics, which show that the case-mix factors are better at predicting perinatal mortality (\( c = 0.853, r^2 = 0.268 \)) than maternal readmissions (\( c = 0.605, r^2 = 0.014 \)).
### Table 49: Correlation between adjusted and unadjusted rates, and model fit statistics

<table>
<thead>
<tr>
<th>Label</th>
<th>Indicator</th>
<th>Correlation coefficient</th>
<th>Model fit statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Pearson</td>
<td>Spearman’s rank</td>
</tr>
<tr>
<td>a</td>
<td>Early caesarean section</td>
<td>0.849**</td>
<td>0.935**</td>
</tr>
<tr>
<td>b</td>
<td>Perineal tear</td>
<td>0.903**</td>
<td>0.851**</td>
</tr>
<tr>
<td>c</td>
<td>Puerperal infection</td>
<td>0.868**</td>
<td>0.971**</td>
</tr>
<tr>
<td>d</td>
<td>30-day maternal emergency readmissions</td>
<td>0.969**</td>
<td>0.972**</td>
</tr>
<tr>
<td>e</td>
<td>In-hospital perinatal mortality</td>
<td>0.540**</td>
<td>0.576**</td>
</tr>
<tr>
<td>f</td>
<td>Injury to neonate</td>
<td>0.948**</td>
<td>0.965**</td>
</tr>
<tr>
<td>g</td>
<td>Selected neonatal infections</td>
<td>0.923**</td>
<td>0.913**</td>
</tr>
<tr>
<td>h</td>
<td>28-day neonatal emergency readmissions</td>
<td>0.936**</td>
<td>0.945**</td>
</tr>
</tbody>
</table>

** Notes:**
- ** p < 0.0001;
- r²: the max-rescaled r-square score reported from PROC LOGISTIC.

#### 9.5.5.2. Coding depth (diagnoses)

Between sites the average number of distinct diagnosis codes used (excluding those relating to perineal tear) per admission ranges from 2.8 to 6.6. Similarly, between sites the average number of distinct diagnosis codes used (excluding those relating to injury to neonate) per admission ranges from 1.1 to 4.4. Compared to the standard logistic model, including a multi-level model with a variable to account for coding practice made little difference to trusts performance in the perineal tear measure (Pearson correlation coefficient, r = 0.986) although there was a more substantial effect for the injury to neonate measure (r = 0.794) (Table 50).
Table 50: Effect of including coding practice in case-mix model

<table>
<thead>
<tr>
<th>Indicator</th>
<th>Correlation coefficient</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Pearson</td>
<td>Spearman’s rank</td>
</tr>
<tr>
<td><strong>Perineal tear</strong> (v. adjusted rates ‘standard’ model)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Multi-level model (including site identifier)</td>
<td>0.991**</td>
<td>0.997**</td>
<td></td>
</tr>
<tr>
<td>Multi-level model, including site coding variable</td>
<td>0.986**</td>
<td>0.983**</td>
<td></td>
</tr>
<tr>
<td><strong>Injury to neonate</strong> (v. adjusted rates ‘standard’ model)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Multi-level model (including site identifier)</td>
<td>0.992**</td>
<td>0.989**</td>
<td></td>
</tr>
<tr>
<td>Multi-level model, including site coding variable</td>
<td>0.794**</td>
<td>0.809**</td>
<td></td>
</tr>
</tbody>
</table>

Note: ** p < 0.0001

For the perineal tear measure, at this site-level there was no correlation (r = -0.029, p = 0.666) between coding depth (measured as average number of unique diagnosis codes in an admission) and performance. Of the 38 sites identified at the 99.8% level in the original regression, all (100.0%) were again flagged as outliers (and in the same direction) at this significance level when coding practice was included in the regression.

For the injury to neonate measure, at this site-level there was a significant correlation (r = 0.260, p = 0.0002) between coding depth (measured as average number of unique diagnosis codes in an admission) and performance. Of the 114 sites identified at the 99.8% level in the original regression, 77 (67.5%) were again flagged as outliers (and in the same direction) at this significance level when coding practice was included in the regression. In addition, three of the original outliers (2.6%) were flagged as outliers in the opposite direction.

9.5.5.3. Coding depth (characteristics)

As shown above (Table 44, p206), the maternal characteristics of the cohort treated by sites with good coding depth on birth-weight and gestational age, were statistically different to those at poor coding sites, although the actual differences are small.

Across poor coders, 44 of 71 sites were outliers at the 99.8% level for the injury measure whereas 70 of the 121 good coders were. The distribution of performance (Higher 0.001, High 0.025, Within control, Low 0.025, Low 0.001) was not statistically different between good and bad coders (p = 0.438).
Across poor coders, 8 of 55 were outliers for perineal tear measure whereas 30 of 161 of good coders were. The distribution of performance (‘Higher 0.001’, ‘High 0.025’, ‘Within control’, ‘Low 0.025’, ‘Low 0.001’) was not statistically different between good and bad coders (p = 0.485).

9.6. Discussion

9.6.1. Summary of findings
The results show the potential for using hospital administrative data, meeting the Government’s call to develop measures of performance and, moreover, to highlight potentially significant variations in the quality and safety across the delivery admission. Eight measures of quality and safety, covering the perinatal care pathway, were applied to English hospital administrative data identifying almost 700,000 births. All eight indicators identified trusts and units with statistically outlying performance at the 99.8% level. Further analysis suggested that differences in coding practice appeared to only partially explain the variation. The analysis reiterated the high impact of maternity care with large numbers of potential adverse events.

9.6.2. Amenable: scope for improvement
For most of the measures, there is no clinical consensus or guidelines for what actual levels are acceptable. In these cases, hospitals should strive to operate at the level of the best performers, after accounting for random variation and variation in case-mix. Indeed, regulators and guideline-setters could use this analysis to suggest levels of performance that are attainable. The exception to this rule is on caesarean sections prior to 39 weeks, for which the extant guidelines state that “the risk of respiratory morbidity is increased in babies born by CS before labour, but this risk decreases significantly after 39 weeks [and] therefore planned CS should not routinely be carried out before 39 weeks” (p18). As such, the current rate of around a third of elective caesarean sections being in contravention to these guidelines seems unacceptable.

9.6.3. Precision
By using funnel plots and control limits, I was able to account for random variation; if the deviation were entirely due to random variation, you would only expect to identify around one outlying maternity unit every 2 measures (given there are 229 units, with an even lower likelihood for across the lower number of trusts). Further to chance, differences in performance may be due to case mix, how the data were collected, or quality of care. Given the ambition was to compare providers on the basis of the last of these factors, the previous two also need to be accounted for and are, therefore, discussed below.
The panels of funnel plots above look like there is less over-dispersion when analysed at a site level. Indeed, there seems to be a smaller proportion of sites than trusts identified as outliers with, for instance, 33 trusts highlighted as outliers at the 99.8% confident level and only 32 sites. However, this may be explained to some extent by the large number of very small units (those clustered on the left of the funnel plots) where numbers are insufficient to give appropriate power to identify variations in performance.

9.6.4. Minimum bias

9.6.4.1. Adjusting for case-mix

The influence of maternal and foetal risk factors on outcomes is well-recognised. I accounted for case-mix using mother- or neonate-level logistic regression to calculate the expected number of events for outcome measures. However at an individual-level some significant case-mix factors, such as maternal obesity and smoking, are not recorded within the data and, therefore, some of the variation may still be caused by differences in case-mix. If we hypothesise that there are small areas of extreme case-mix but, at a larger geographical level, case-mix is more homogenous coupled with units and trusts predominantly treating patients from their local area then it holds that any bias from this potential 'categorisation error' is less likely to have a substantial effect within the trust-level analysis given that at least 1,000 deliveries occurred at each provider, in comparison to the unit-level results whereby one unit only had 10 deliveries.

I report, in detail, the model fit statistics and effect of the case-mix adjustment of hospitals' performance. This showed that the case-mix models were better at explaining performance in the perinatal mortality score to the readmission measures. This information is not for making conclusion on the appropriateness of the case-mix adjustment – since it is not possible to disaggregated the effect of unmeasured differences in case-mix to differences in quality of care provided – but should be considered when interpreting the results. One specific issues relating to case-mix originates from the difference between obstetrician-led and midwifery-led care. Midwifery-led units are recommended only for low-risk mothers and, therefore, obstetric units are more likely to a higher proportion of high-risk cases.

9.6.4.2. Accounting for variation in coding practice

The variation in performance due to how data were collected is harder to disaggregate since some of the possible variation in coding practice, such as depth of diagnosis coding, might mimic differences in case-mix. However, I showed that two key coding issues that were central to the assumptions for identifying complications and adjusting for case-mix explained only a minority of the differences in
performance against the measures. Previous research has raised questions about differences in performance between hospitals with good and poor data recording – for instance, a higher postoperative mortality rate following resection of colorectal cancer was found in trusts that do not voluntarily report data to a national audit – the findings here suggest that outliers were in both good and bad recording trusts.

9.6.5. Errors in underlying data

Records of births outside of NHS hospitals are significantly affected by data quality issues and do not contain information for fields such as admission/discharge date (admidate/disdate) or patient classification (classpat). Historically there has been poorer coverage of home deliveries, and for this reason, only records for births in an NHS setting are included in the HES Annual Maternity publication. Yet there are wide geographical and demographic variations in the proportion of home births, which could cause bias if creating performance indicators without these home births. Moreover, some of this missing data can be estimated from the date of birth and, as such, the research presented here did not actively exclude births outside hospitals.

The additional fields have their own data quality issues and the coverage is not as complete as for the rest of HES data. For instance, a research project suggested that key data items were missing in over 20% of records overall, with some NHS trusts not submitting usable maternity tail data for any deliveries. However, I investigated these issues by disaggregating the analysis between good and poor coding hospitals, as reported above.

9.6.5.1. Case-ascertainment strategy

If the total number of cases attained from the algorithm applied to the HES extract was either substantially higher or lower than that derived from the national birth register then this would suggest either poor sensitivity or specificity of the case-ascertainment strategy. Reassuringly, however, the number of maternities and births from the extract were similar to the register. The numbers of maternities and births from the HES data closely matches the register data – differing by only 0.3 and 0.2%, respectively, in 2011 – and has improved over time (Table 51).
Table 51. Comparison of numbers of maternities and births from the HES extract and ONS births register

<table>
<thead>
<tr>
<th></th>
<th>Maternities</th>
<th></th>
<th>Live and still births</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Register</td>
<td>HES Extract</td>
<td>Variation</td>
<td>Register</td>
</tr>
<tr>
<td>2008</td>
<td>665,779</td>
<td>643,658</td>
<td>-3.3%</td>
<td>676,236</td>
</tr>
<tr>
<td>2009</td>
<td>645,802</td>
<td>640,957</td>
<td>-0.8%</td>
<td>656,880</td>
</tr>
<tr>
<td>2010</td>
<td>662,752</td>
<td>661,162</td>
<td>-0.2%</td>
<td>673,641</td>
</tr>
<tr>
<td>2011</td>
<td>664,468</td>
<td>666,352</td>
<td>+0.3%</td>
<td>675,646</td>
</tr>
</tbody>
</table>

Notes: ONS data from Characteristics of Birth 1

9.6.6. Face and content validity

Whilst previous studies have used hospital administrative data to measure the performance of aspects of obstetric care, none have brought together such a broad range of measures covering both maternal and neonatal care and, instead, have tended to focus on a single area such as perineal tears. A combination of process and outcome measures (both short- and longer-term) was used in this study, so benefiting from the advantages of process measures (which tend to be more sensitive to differences in the quality of care and offer a clear action for improvement) and outcome measures (greater intrinsic interest, high face validity, and can reflect all aspects of care, including those that are otherwise difficult to measure such as technical expertise and operator skill). Face validity can be derived from the fact that the indicators appeared in existing literature or, for one of the measures, clinical guidelines.

The overall level of quality of care at a provider-level cannot be deduced from the results of only a small set of indicators; however almost half of studies within the literature review applied only one indicator. This study included eight indicators covering many aspects of the care provided during the delivery admission and, therefore, is likely to have better content validity (extent to which a measure represents all facets of the subject being investigated) over these previous more limited evaluations.

9.6.7. Criterion validity

As well as these considerations on whether the indicators’ algorithms are functioning as intended, further assurance needs to be taken about whether the results of the indicators reflect actual performance. The most common process for validating the results is to compare against a gold standard. Previous research has focused on the perineal tear (obstetric trauma) indicator, including: the California Obstetric Validity Study which found that the indicator had sensitivity of 90% (95% CI, 82-96%) and PPV of 90-95%; a study using a clinical research dataset as comparison which
suggested sensitivity of 77% (72-81%) and PPV of approximately 93%;\textsuperscript{494} and, within England, research suggested a PPV of 85% based on a review of 995 case notes.\textsuperscript{112}

9.6.8. Construct validity

In the absence of a gold standard dataset to compare results, some assurance on validity can be derived by comparing with other indicators of performance. I investigated whether there were associations between trusts’ and units’ performance across the different measures. One might hypothesise there would be correlation between good performance across all the indicators with, for example, good management of maternity care likely to affect all the quality and safety measures to some extent. More specifically, you would expect a hospital with good hygiene levels to have good performance against both the maternal and neonatal infection measure. Ideally there would be a range of outcomes and process measures focusing on the same event within the delivery to support this validation; however, the literature review revealed only limited process measures and these were not directly linked to the outcome measures.

There were 6 statistically significant correlations at the 95% confidence level at the trust-level, and 7 at the unit-level. However, only 2 of these were common to both trust- and unit-level, which could suggest that there are different unit and trust-level effects that affect performance.

At the trust-level, the only correlation significant at the 99.8% level was a positive correlation ($r = 0.198$) in the direction of good performance in both maternal readmissions and perineal tears, which might be expected as poor care during the delivery, as indicated by obstetric trauma, is likely to result in readmissions after discharge.

As hypothesised, at the unit-level there is a strong positive association between neonatal and maternal infections ($r = 0.259$, $p < 0.002$). Across the units and at the 99.8% confidence level, there was also positive statistically significant correlations between perinatal mortality and both neonatal infections ($r = 0.355$) and maternal readmissions ($0.218$).

There were further statistically significant correlations at the 95% confidence level. Some of these (such as injury to neonate and puerperal infection at both trust and unit levels) showed a positive association between complication rates, as expected. However, at this significance level, there were also unexpected negative associations (neonatal readmissions with perineal tear at the site and trust levels; and at a trust-level only for early caesarean sections and injury to neonate and between neonatal infections and perineal tear). It is unclear why these correlations are found; however, across the 56 comparisons we might expect around 3 significant correlations by chance alone at this level which might provide the explanation.
9.6.9. Unit of analysis

9.6.9.1. Provider identification

While I was able to identify the trust-level ‘NHS code’ for every birth and maternity, the attribution to specific maternity unit was more complex. The matching of fields for identifying the unit of provider identified, for 2011-12, 224 distinct geographic locations – either obstetric unit, obstetric unit and alongside freestanding midwifery unit, or freestanding midwifery unit – and 5 providers where I was unable to disaggregate the distinct locations, although for 2 of these providers the separate units were of the same type. In comparison a national audit of maternity surveys identified 229 distinct geographical locations, as at June 2013. The closeness of these numbers suggests that the HES fields can be used to identify the vast majority of maternity units.

Performance might be affected by both trust-level factors – e.g. overall staffing levels given that midwives are often used flexibly between units – and unit factors – e.g. suitability of the estate. As discussed above, analysis at the trust level will better control for unmeasured differences in case-mix. However this more aggregated level might not be sensitive to some of the local differences in quality. As such, there seems to be value to analyse the results at both levels.

Whereas the report by the Royal College of Obstetricians and Gynaecologists was limited to those units with greater than 1000 deliveries, I included the full set here. This inclusion criterion by the Royal College seems unnecessary as they too apply a statistical approach that accounts for the random variation which is likely to have a greater effect on smaller units.

9.6.10. Strengths of this study to other measurement efforts

The Confidential Inquiry into maternal mortality has been commonly cited and linked to many of the national standards that had been used in the scheme for providing litigation cover from trusts in England (Clinical Negligence Scheme for Trusts). However, the old confidential inquiry was also widely criticised for being “anecdotal”, expensive and without evidence of benefit. The RCOG have made advances in measuring performance, however, the set of indicators used considered only maternal process and outcomes of care, and did not compare sites performance across the different indicators. Indeed, the RCOG recommended that “there is also a need to better understand the relationships between different process and outcome indicators. Priority areas should be neonatal outcomes and measures of user experience.”

---

8 County Durham and Darlington (RXP) includes two obstetric units (RXP00 and RXPBA). Doncaster and Bassetlaw (RPS) includes two obstetric units.
HES has the advantage of being: longitudinal; timely; covering all hospital admissions; and being relatively cheap, costing £1 per record to collect compared with around £10 - £60 per record for clinical registers.\(^8\) The introduction of a National Maternity Dataset for England, mandated from April 2014, should provide a rich and accurate source of information on the care of pregnant women.\(^3\)\(^7\)\(^5\)

9.6.11. Conclusion

This study shows that hospital administrative data have the potential for benchmarking the standard of hospitals performance over the care provided in the delivery admission, in a timely way and without increasing the bureaucratic burden on hospitals. While the study benefits from drawing together both neonatal and maternal indicators being able to link mother and baby records – for instance to create a compound measure of ‘no complication form delivery’ – would give a more complete picture of performance.\(^3\)\(^7\)\(^5\)

As has been previously noted above and reiterated by others, “the data on which our analyses are based are not as accurate as we would have liked; however, until these data are used to provide information to allow for meaningful benchmarking, encouraging clinicians to take ownership of their own hospital data and attempts to drive up quality will be difficult.” (p. viii)\(^3\)\(^7\)\(^5\) There are a number of reasons for the coverage and data quality issues, such as different capabilities of local systems to record the additional birth details.\(^4\)\(^9\)\(^7\) Indeed, a recent audit found that nearly a fifth of maternity units’ information systems were either heavily reliant on paper-based notes or not linked to the patient administration system, which will negatively affect the ability to provide accurate, timely data.\(^1\)\(^5\)\(^2\) However, the government recently stated that they were ensuring that all maternity units would be required to have appropriate IT.\(^4\)\(^9\)\(^8\)

Whilst neither the lack of credible gold standard nor evidence of bias from coding practice render the indicators redundant, it does temper the extent to which they could be used, such as for regulating hospitals. This study shows that HES provides the facility to record some key process and outcome measures across key aspects of maternity care in a cheap and timely manner. These results could be linked to structural measures, such as unit-level data on staffing levels, to investigate their effectiveness.\(^3\)\(^7\)\(^5\) With suitable modification and further development, these data could also be used routinely as currently measurement efforts.
Chapter 10.

Obstetrics temporal trends
PART III: Obstetrics

Overview

Context

The previous chapter showed how obstetric indicators can be used to highlight variations in the quality and safety of care and demonstrated that, while the findings are relatively robust to differences in coding practice, this does represent a limitation. This chapter shows how these limitations can be overcome through different application of the indicators. The literature review revealed that less than a quarter of studies on obstetric care had used the longitudinal nature of administrative data, and none had comprehensively assessed weekend care.

Methods

Application of indicators covering key aspects of the delivery admission to investigate changes in performance over time and association between day of delivery and adverse events. Results were adjusted for case-mix using logistic regression.

Findings

The longitudinal analysis highlighted both seasonal and longitudinal trends. The day of the week analysis showed performance across four of the seven measures was significantly lower at weekends. In particularly, the perinatal mortality rate was 7.3 per 1,000 at weekends, some 0.9 per 1,000 higher than at weekdays (adjusted odds ratio 1.071; 95% Confidence Interval 1.019 – 1.126).

What this chapter adds

This chapter presents the most comprehensive evaluation of the “weekend effect” in obstetric care. The study design mitigates for common limitations cited on the use of such indicators and so is able to produce robust results about the extent of this important issue.
Chapter 10: Obstetrics temporal trends

10.1. Background

10.1.1. Existing obstetric studies on administrative data

10.1.1.1. Topics, designs and results

The literature review (Chapter 8) revealed that few studies have taken advantage of the longitudinal nature of administrative data to evaluate temporal trends in the quality and safety of obstetrics care. Where studies have existed, the temporal trends investigated have included longitudinal annual trends, seasonal trends (a so called “July” effect) and association with day of admission (Table 52).

Only one of these studies was based on English data, and this evaluation included only timing of caesarean sections as a measure.\textsuperscript{411} The treatment of changes in case-mix and patient characteristics varied amongst this small set of studies with, for instance, Callaghan and colleagues (2012) not making any risk-adjustment to the evaluation in changes in maternal mortality and severe morbidity.\textsuperscript{442} Most of these studies fail to stand up to the criticism of primarily focusing on a single outcome and therefore not capturing wider aspects of the quality and safety of care.

The study designs and aims were heterogeneous and so a formal meta-analysis is not possible but, in summary, they present a mixed picture on longitudinal trends with some countries showing improvement in certain indicators and worsening performance in others. Neither the study of the “July” effect\textsuperscript{444} nor “weekend effect”\textsuperscript{445} showed statistically significant variations in quality and safety.

10.1.2. Other literature

10.1.2.1. Longitudinal and temporal trends

A recent national audit suggested that there had been improvement across some key aspects of maternity care over time, such as early access and consultant presence.\textsuperscript{152}

10.1.2.2. Association between performance and day of the week

Previous studies, across a range of countries, have identified higher mortality in patients admitted on weekends across a range of medical conditions; a phenomenon termed the ‘weekend effect’.\textsuperscript{253,350-353,499} This calls into question the idea that quality of care is equal irrespective of when you present at hospital. However, not all studies have identified an association between poor outcomes and out-of-hours periods.\textsuperscript{500-502}
In 1978, Alison MacFarlane published a paper showing a seven-day cycle in birth numbers across
England (and Wales) and that perinatal mortality was higher among babies born at weekends.\textsuperscript{503} Similar studies in the 1970s and 1980s found this phenomenon in other developed countries.\textsuperscript{504-506} The delivery of obstetrics care has changed dramatically since this time; however, where the
“weekend effect” has been evaluated it has predominantly been based on mortality as the sole
indicator. Indeed, in setting out key challenges to address in obstetrics care, a paper by the World
Health Organisation (WHO) highlighted ineffective referral to and inadequate availability of 24-hour
quality services to emergency obstetric care services.\textsuperscript{367}

10.2. Objectives

In this chapter I investigate the potential to use the longitudinal nature of administrative data to
identify important shortcomings in quality and safety in a way that overcomes some of the known
limitations to such indicators.
<table>
<thead>
<tr>
<th>Study</th>
<th>Cohort</th>
<th>Country</th>
<th>Variable</th>
<th>Measures</th>
<th>Risk-adjustment</th>
<th>Main results</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cameron et al (2006)</td>
<td>52,151 women with PPH (1994 – 2002)</td>
<td>Australia</td>
<td>Year</td>
<td>Primary measures: perinatal PPH and readmissions from PPH</td>
<td>Age and mode of delivery Adjusted for under-reporting</td>
<td>From 1994-2002, PPH increased from 8.3 to 10.7% of deliveries. Hospital readmission for PPH declined from 1.2% to 0.9%.</td>
</tr>
<tr>
<td>Gurol-Urgancil et al (2011)</td>
<td>507,410 low risk singletons (2000 – 09)</td>
<td>England</td>
<td>Year</td>
<td>Elective CS after 39 weeks</td>
<td>None – excluded women who had elective CS before 24 weeks or contra-indication for delaying an elective CS</td>
<td>Elective CS deliveries after 39 weeks steadily increased from 39% in 2000/01 to 63% in 2008/09. Average annual decrease in events: birth trauma -17.8%; obstetric trauma without -5.7%; obstetric trauma with -4.1%.</td>
</tr>
<tr>
<td>Roberts et al (2012)</td>
<td>500,603 women (1999 – 2004)</td>
<td>Australia</td>
<td>Year</td>
<td>Maternal morbidity outcome indicator (MMOI) PPH</td>
<td>Age, parity, mode of delivery, smoking, delivery hospital, multiple pregnancy, previous CS, and other clinical factors</td>
<td>MMOI increased annually by an average 3.8% (95% CI: 2.3-5.3%).</td>
</tr>
</tbody>
</table>
10.3. Methods

10.3.1. General methodological considerations

10.3.1.1. Indicators

The indicators used for this part of the study are similar to those used in the previous chapter (paragraph 9.3.3.2, p198) and were identified in the literature review (Chapter 8). In summary, the indicators used in this chapter are set out in the below (Table 53), although some amendments were used in the weekend analysis, with the rationale for this discussed below.

Table 53. Obstetric indicators

<table>
<thead>
<tr>
<th>Delivery</th>
<th>Maternal</th>
<th>Neonatal</th>
</tr>
</thead>
<tbody>
<tr>
<td>a. Early (pre-39 week) caesarean section</td>
<td>b. Perineal tear</td>
<td>e. In-hospital perinatal mortality</td>
</tr>
<tr>
<td>c. Puerperal infection</td>
<td>f. Injury to neonate</td>
<td></td>
</tr>
<tr>
<td>d. 30-day emergency readmission rates</td>
<td>g. Selected neonatal infections</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>h. 28-day emergency readmission rates</td>
</tr>
</tbody>
</table>

10.3.1.2. Data extract

The details of deliveries from 1 April 2009 to 31 March 2012 were extracted from the HES database. The indicator definitions were applied to the extract to obtain denominator and numerators, categorised by either month or day of admission (for maternal indicators) or birth (for neonatal records). The weekend analysis was restricted to 1 April 2010 to 31 March 2012 to reduce any bias from poorer coding in the preceding year.

10.3.1.3. Case-mix adjustment

Variables expected to influence the outcome of the association between the quality and safety indicators and day of admission/birth were extracted, including gestational age, birth-weight, and maternal age. A full set of case-mix factors and the method for risk-adjustment is covered in paragraph 9.4.2, p202.
10.3.1.4. Data analysis and software

National performance across the measures was calculated, disaggregated by either month or day of admission/birth. Multiple logistic regressions were used to adjust to the effect of the covariates. I did not adjust for the clustering of patients within hospital as the hospital-level effects were found to be small (as shown in Chapter 9) except when including site level data (such as hours of consultant presence) as described explicitly below. Results were also displayed by plotting – using either April 2009 (or Tuesday) as a reference – odds ratios and 95% confidence intervals, by month (or day) of delivery.

Analyses were carried out using SAS Version 9.2, using the PROC LOGISTIC and PROC GLIMMIX procedure for regression analyses.

10.3.2. Additional ‘day of the week’ methodological considerations

10.3.2.1. Defining ‘weekend’

Where results are presented as weekend versus weekday performance, the former was defined as the period from midnight on Friday to midnight on Sunday, with all other times defined as weekdays (the time of admission is not captured in HES).

10.3.2.2. Estimating impact

Using regression analysis on just Tuesdays, probabilities of in-hospital perinatal death and puerperal infections (maternal) were derived. By matching these probabilities for each Tuesday admission, based on the mother’s or neonate’s characteristics, indirectly standardised estimates for the outcomes as if those non-Tuesday cases had had similar rates as their Tuesday counterparts were calculated, with these particular results reported in the discussion. I used Tuesday (rather than Monday, as for the stroke analysis) as the reference since the neonatal indicators use date of birth to assign the day even through the outcome might be affected by the standard of care given to the mother during labour and, as such, Monday’s performance might be affected by quality of care over the weekend for non-same-day births.

10.3.2.3. Amendments to indicators

The amendments made to these indicators, for the purpose of this day-of-the-week study, were to:

- use 3-day readmission rates for both maternal and neonatal indicators in keeping with the suggestion that it is a more appropriate timeframe to evaluate the association between day of admission and mortality. 354,499
- drop the indicator for early caesarean sections as such deliveries are not scheduled during out-of-hours periods.

10.3.2.4. Consultant presence

Data on hours of consultant presence on the labour ward were collected obtained from two previous audits, by the National Audit Office\textsuperscript{152} and Royal College of Obstetricians and Gynaecologists (RCOG).\textsuperscript{375} Data were linked using unit names with 100% completeness. A binary indicator of compliance with was created based on recommendations RCOG as set out in the table below (Table 54). The analysis was restricted to obstetric units and sites where data were reported for the obstetric and alongside midwifery unit combined.

Table 54. Recommended levels of consultant presence

<table>
<thead>
<tr>
<th>Births per year</th>
<th>Recommended minimum consultant presence per week</th>
</tr>
</thead>
<tbody>
<tr>
<td>2,500 – 4,000</td>
<td>60 hours</td>
</tr>
<tr>
<td>4,000 – 5,000</td>
<td>98 hours</td>
</tr>
<tr>
<td>5,000 +</td>
<td>168 hours</td>
</tr>
</tbody>
</table>

Note: Source table quoted in National Audit Office report.\textsuperscript{152}

10.3.2.5. Association between levels of activity and performance

The level of activity, $A$, was defined using the following equation:

$$A_{u,y} = \begin{cases} 
0 & \text{if } \frac{n_{u,d,y}}{\sum_{d} n_{u,d,y}} \leq 0.8 \\
1 & \text{if } \frac{n_{u,d,y}}{\sum_{d} n_{u,d,y}} > 0.2 
\end{cases}$$

where $u = \text{unit } u$;

$d = \text{day } \epsilon \{1 \text{ April}, ..., 31 \text{ March}\}$;

$y = \text{year } \epsilon \{2011, 2012\}$
10.4. Results

10.4.1. Longitudinal analysis

10.4.1.1. Overview of results

Between April 2009 and March 2012, I identified around 4 million births and maternities. The most common adverse event was early caesarean section (37.50%) with the least common being in-hospital perinatal mortality (0.67%) (Table 55).

Table 55. Number of births and maternities and complication rates

<table>
<thead>
<tr>
<th>Measure</th>
<th>Total</th>
<th>2009-10</th>
<th>2010-11</th>
<th>2011-12</th>
</tr>
</thead>
<tbody>
<tr>
<td>Maternities</td>
<td>1,979,360</td>
<td>n = 646,525</td>
<td>n = 663,218</td>
<td>n = 669,617</td>
</tr>
<tr>
<td>Births</td>
<td>2,008,144</td>
<td>n = 658,545</td>
<td>n = 670,814</td>
<td>n = 678,785</td>
</tr>
<tr>
<td>a Early caesarean section</td>
<td>49,074</td>
<td>37.50%</td>
<td>39.5%</td>
<td>38.0%</td>
</tr>
<tr>
<td>b Perineal tear</td>
<td>43,890</td>
<td>2.94%</td>
<td>2.73%</td>
<td>3.01%</td>
</tr>
<tr>
<td>c Puerperal infection</td>
<td>15,521</td>
<td>0.78%</td>
<td>0.68%</td>
<td>0.74%</td>
</tr>
<tr>
<td>d 30-day maternal emergency readmissions</td>
<td>18,423</td>
<td>1.02%</td>
<td>1.04%</td>
<td>1.04%</td>
</tr>
<tr>
<td>e In-hospital perinatal mortality</td>
<td>13,402</td>
<td>0.67%</td>
<td>0.67%</td>
<td>0.66%</td>
</tr>
<tr>
<td>f Injury to neonate</td>
<td>26,005</td>
<td>1.40%</td>
<td>1.26%</td>
<td>1.41%</td>
</tr>
<tr>
<td>g Selected neonatal infections</td>
<td>38,156</td>
<td>1.91%</td>
<td>1.74%</td>
<td>1.90%</td>
</tr>
<tr>
<td>h 28-day neonatal emergency readmissions</td>
<td>106,799</td>
<td>5.85%</td>
<td>5.53%</td>
<td>5.98%</td>
</tr>
</tbody>
</table>

Note: Percentage of cases meeting inclusion criteria. Readmission figures exclude final month of data year to allow comparability over years.

In the first instance, I plotted the rates per month to better understand what trends might be present (Figure 29). Some of the measures, such as early caesarean sections showed a relatively consistent trend over time and others, for example neonatal infections, showed a seasonal trend.
Figure 29: Trends in obstetric indicators, by month

a. Early caesarean section

b. Perineal tear

c. Puerperal infection

d. 30-day maternal emergency readmissions

e. In-hospital perinatal mortality

f. Injury to neonate

g. Selected neonatal infections

h. 28-day neonatal emergency readmissions

Note: Data points represent odds ratios, with April 2007 used as a reference (1.00); vertical ranges, 95% confidence intervals.
10.4.1.2. Analysis of longitudinal trends and seasonal effect

The more detail analysis revealed that all indicators showed a statistically significant (at the 95% confidence level) linear longitudinal trend, with six of the eight measures showing a worsening in performance. There were 8 instances of months with statistically different performance to the reference (April) at the 99.8% level, covering early caesarean sections, neonatal infections and neonatal readmissions (Table 56).

Table 56. Longitudinal trend and seasonal effect

<table>
<thead>
<tr>
<th></th>
<th>a. Early caesarean section</th>
<th>b. Perineal tear</th>
<th>c. Puerperal infection</th>
<th>d. 30-day maternal emergency readmissions</th>
<th>e. In-hospital perinatal mortality</th>
<th>f. Injury to neonate</th>
<th>g. Selected neonatal infections</th>
<th>h. 28-day neonatal emergency readmissions</th>
<th>No. indicators significantly better</th>
<th>No. indicators significantly worse</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Longitudinal trend</strong></td>
<td>0.990**</td>
<td>1.004**</td>
<td>1.014**</td>
<td>0.998*</td>
<td>1.003**</td>
<td>1.007**</td>
<td>1.009**</td>
<td>1.003**</td>
<td>2</td>
<td>6</td>
</tr>
<tr>
<td>January</td>
<td>0.975</td>
<td>1.027</td>
<td>1.004</td>
<td>0.953</td>
<td>0.942</td>
<td>1.036</td>
<td>0.899**</td>
<td>1.049**</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>February</td>
<td>1.001</td>
<td>1.032</td>
<td>1.067</td>
<td>0.985</td>
<td>0.927</td>
<td>1.026</td>
<td>0.920*</td>
<td>1.041*</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>March</td>
<td>0.975</td>
<td>1.033</td>
<td>0.977</td>
<td>0.987</td>
<td>0.889*</td>
<td>1.002</td>
<td>0.837**</td>
<td>1.025</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>April</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>May</td>
<td>0.957</td>
<td>1.033</td>
<td>0.978</td>
<td>1.063</td>
<td>1.011</td>
<td>1.076*</td>
<td>1.039</td>
<td>1.006</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>June</td>
<td>0.948</td>
<td>0.957</td>
<td>1.049</td>
<td>0.99</td>
<td>0.968</td>
<td>1.092*</td>
<td>1.002</td>
<td>1.038*</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>July</td>
<td>0.982</td>
<td>0.946*</td>
<td>1.090*</td>
<td>1.039</td>
<td>1.013</td>
<td>1.089*</td>
<td>1.019</td>
<td>1.003</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>August</td>
<td>1.011</td>
<td>0.940*</td>
<td>1.073</td>
<td>1.026</td>
<td>0.957</td>
<td>1.074*</td>
<td>1.064*</td>
<td>0.975</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>September</td>
<td>0.884**</td>
<td>0.955</td>
<td>1.06</td>
<td>1.024</td>
<td>1.049</td>
<td>1.018</td>
<td>0.98</td>
<td>1.023</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>October</td>
<td>1.003</td>
<td>1.019</td>
<td>1.017</td>
<td>0.997</td>
<td>0.948</td>
<td>1.038</td>
<td>1.008</td>
<td>1.062**</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>November</td>
<td>0.982</td>
<td>1.043</td>
<td>1.001</td>
<td>0.932</td>
<td>0.901*</td>
<td>1.047</td>
<td>0.957</td>
<td>1.141**</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>December</td>
<td>1.125**</td>
<td>1.01</td>
<td>1.004</td>
<td>0.914*</td>
<td>0.939</td>
<td>0.99</td>
<td>0.934*</td>
<td>1.123**</td>
<td>2</td>
<td>2</td>
</tr>
</tbody>
</table>

Note: 1. Longitudinal trend measured as linear, with positive odds ratio suggesting increasing complication rates. Seasonal trend measured as performance in month in comparison to April.

2: Correlation significant at the 95% level marked with ‘**’; those at 99.8% level with ‘***’
10.4.2. Association with day of the week

10.4.2.1. Descriptive overview

The distribution of births is not even across the days of the week (Figure 30). The most common day for giving birth is a Thursday (15% of births and maternities) and the least common is Sunday (12%). Much of the difference is explained by number of elective caesarean sections performed during weekdays.

Figure 30. Proportion of births and maternities and breakdown in number of elective caesarean sections, by day of week

Table 57 gives a description of the characteristics of patients included in the study.

Table 57. Characteristics of the study population

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Categories</th>
<th>Weekdays</th>
<th>Weekends</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Delivery method</td>
<td>Spontaneous vertex</td>
<td>59.7</td>
<td>68.6</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td></td>
<td>Spontaneous, other cephalic</td>
<td>0.5</td>
<td>0.5</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Low forceps, non-breach</td>
<td>3.9</td>
<td>4.6</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Other forceps, non-breach</td>
<td>2.1</td>
<td>2.5</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Ventouse, vacuum extraction</td>
<td>6.1</td>
<td>6.9</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Breech</td>
<td>0.4</td>
<td>0.4</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Breech extraction not otherwise specified</td>
<td>0.1</td>
<td>0.1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Elective caesarean</td>
<td>12.7</td>
<td>1.5</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Emergency caesarean</td>
<td>14.6</td>
<td>14.9</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Other than those specified above</td>
<td>0.1</td>
<td>0.1</td>
<td></td>
</tr>
<tr>
<td>Maternal age</td>
<td>&lt;19</td>
<td>5.1</td>
<td>5.8</td>
<td></td>
</tr>
<tr>
<td>Ethnicity</td>
<td>White</td>
<td>Asian</td>
<td>Other (inc. Chinese)</td>
<td>Black (inc. Black British)Afro-Caribbean</td>
</tr>
<tr>
<td>-----------------------------------</td>
<td>----------</td>
<td>---------</td>
<td>---------------------</td>
<td>----------------------------------------</td>
</tr>
<tr>
<td></td>
<td>72.3</td>
<td>10.3</td>
<td>3.3</td>
<td>5.1</td>
</tr>
<tr>
<td>Level of deprivations</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>15.2</td>
<td>14.9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>‘6’</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Parity</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Primiparous</td>
<td>43.5</td>
<td>46.4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Multiparous</td>
<td>56.5</td>
<td>53.6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gestational age</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt; 37</td>
<td>7.6</td>
<td>7.4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>37-39</td>
<td>35.7</td>
<td>30.3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>40-41</td>
<td>39.7</td>
<td>45.2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>&gt;=42</td>
<td>3.6</td>
<td>4.1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unknown</td>
<td>13.5</td>
<td>13.2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Birth-weight</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt;2500g</td>
<td>5.9</td>
<td>5.4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2500-4000g</td>
<td>74.1</td>
<td>74.7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>&gt;4000g</td>
<td>10.1</td>
<td>10.2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unknown</td>
<td>9.9</td>
<td>9.7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Multip</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>1.6</td>
<td>0.9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>No</td>
<td>91.0</td>
<td>91.8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unknown</td>
<td>7.4</td>
<td>7.3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Previous caesarean section</td>
<td>12.8</td>
<td>6.9</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note: Data are for maternities

10.4.2.1. Association between day of delivery and performance

Table 58 shows the results of the association between weekday/weekend admission and performance in the seven measures of quality and safety. There were statistically significant
associations in four of the indicators, all of which were consistent with a lower standard of care at weekends. The largest effects were seen in the higher rates perinatal mortality rates (adjusted odds ratio [OR] 1.07; 95% Confidence Interval 1.02 – 1.13) and puerperal infections (OR 1.06; 1.01 – 1.11).

Table 58. Association between weekday/weekend admission (or birth) and indicators of quality and safety of care

<table>
<thead>
<tr>
<th>Indicator</th>
<th>Weekday admission</th>
<th>Weekend admission</th>
<th>p-value</th>
<th>OR (95% CI) (weekday as reference)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Unadjusted %</td>
<td>Adjusted rate, %</td>
<td>Unadjusted %</td>
<td>Adjusted rate, %</td>
</tr>
<tr>
<td></td>
<td>(number)</td>
<td></td>
<td>(number)</td>
<td></td>
</tr>
<tr>
<td>b Perineal tear</td>
<td>3.03 (22,299)</td>
<td>3.04</td>
<td>3.07 (8,249)</td>
<td>3.05</td>
</tr>
<tr>
<td>c Puerperal infection</td>
<td>0.83 (8,358)</td>
<td>0.82</td>
<td>0.86 (2,770)</td>
<td>0.87</td>
</tr>
<tr>
<td>d 3-day maternal readmissions</td>
<td>0.20 (1,926)</td>
<td>0.20</td>
<td>0.18 (545)</td>
<td>0.18</td>
</tr>
<tr>
<td>e In-hospital perinatal mortality</td>
<td>0.64 (6,481)</td>
<td>0.65</td>
<td>0.73 (2,518)</td>
<td>0.71</td>
</tr>
<tr>
<td>f Injury to neonate</td>
<td>1.43 (13,278)</td>
<td>1.45</td>
<td>1.59 (5,038)</td>
<td>1.53</td>
</tr>
<tr>
<td>g Selected neonatal infections</td>
<td>1.99 (19,900)</td>
<td>1.99</td>
<td>2.01 (6,852)</td>
<td>2.00</td>
</tr>
<tr>
<td>h 3-day neonatal readmissions</td>
<td>1.19 (11,323)</td>
<td>1.18</td>
<td>1.22 (3,976)</td>
<td>1.23</td>
</tr>
</tbody>
</table>

Notes: CI denotes confidence interval.

Over the seven indicators, there were 11 examples of statistically significant differences in the performance of days in comparison to Tuesday. Compared to this reference day, all days bar Wednesday had statistically higher rates of perinatal mortality (Figure 31).
Figure 31: Association between performance and day of admission/birth

Indicator ‘a. Early caesarean sections’ not used in this part of the study

Note: Data points represent odds ratios, with Tuesday used as a reference (1.00); vertical ranges, 95% confidence intervals.
The weekend effect varied by delivery method (Table 59).

### Table 59 Perinatal rates by delivery rates for common weekend delivery methods

<table>
<thead>
<tr>
<th>Delivery method</th>
<th>Unadjusted % (number)</th>
<th>Adjusted rate, %</th>
<th>Unadjusted % (number)</th>
<th>Adjusted rate, %</th>
<th>OR (95% CI) (weekday as reference)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spontaneous vertex</td>
<td>0.57 (484,998)</td>
<td>0.76</td>
<td>0.61 (189,210)</td>
<td>0.80</td>
<td>1.059 (0.983 – 1.140)</td>
</tr>
<tr>
<td>Unknown</td>
<td>1.12 (133,424)</td>
<td>0.64</td>
<td>1.36 (45,527)</td>
<td>0.76</td>
<td>1.202 (1.093 – 1.322)</td>
</tr>
<tr>
<td>Emergency caesarean</td>
<td>0.60 (116,280)</td>
<td>0.40</td>
<td>0.58 (42,513)</td>
<td>0.41</td>
<td>1.035 (0.890 – 1.204)</td>
</tr>
<tr>
<td>Ventouse, vacuum extraction</td>
<td>0.16 (48,887)</td>
<td>0.34</td>
<td>0.11 (19,161)</td>
<td>0.23</td>
<td>0.676 (0.415 – 1.103)</td>
</tr>
<tr>
<td>Spontaneous, other cephalic</td>
<td>0.81 (36,496)</td>
<td>0.96</td>
<td>0.79 (14,094)</td>
<td>0.92</td>
<td>0.946 (0.746 – 1.199)</td>
</tr>
<tr>
<td>Low forceps, non-breach</td>
<td>0.20 (29,509)</td>
<td>0.39</td>
<td>0.17 (11,704)</td>
<td>0.33</td>
<td>0.836 (0.501 – 1.397)</td>
</tr>
</tbody>
</table>

Note: Common weekend delivery methods defined as > 5,000 deliveries per year at weekends

### 10.4.2.1. Association between activity and performance

For sites, high activity times accounted for 9.67% of maternities and 10.09% of births. For trusts, the corresponding figures were 11.10% and 11.37% respectively. The logistic non-hierarchical regression model suggested higher perinatal mortality rates at busy times (Table 60). However, when provider-level effects are added, this raised mortality is no longer statistically significant at the 95% level.
Table 60. Association between business and complication rates, for weekdays April 2010 – March 2012

<table>
<thead>
<tr>
<th>Measure</th>
<th>Trust level¹ (busy vs not busy)</th>
<th>Trust level¹ with trust effects (busy vs not busy)</th>
<th>Unit level with unit level effects² (busy vs not busy)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>OR (95% CI)³</td>
<td>p-value (95% CI)³</td>
<td>OR (95% CI)³ p-value</td>
</tr>
<tr>
<td>Perineal tear</td>
<td>0.986 (0.943 – 1.031)</td>
<td>0.55</td>
<td>0.981 (0.938 - 1.027)</td>
</tr>
<tr>
<td>Puerperal infection</td>
<td>0.999 (0.930 – 1.073)</td>
<td>0.98</td>
<td>0.953 (0.887 - 1.026)</td>
</tr>
<tr>
<td>Maternal readmissions</td>
<td>0.992 (0.929 – 1.060)</td>
<td>0.82</td>
<td>1.028 (0.933 - 1.131)</td>
</tr>
<tr>
<td>In-hospital perinatal mortality</td>
<td>1.101 (1.014 – 1.195)</td>
<td>0.02</td>
<td>1.072 (0.985 - 1.166)</td>
</tr>
<tr>
<td>Injury to neonate</td>
<td>1.022 (0.965 – 1.082)</td>
<td>0.46</td>
<td>0.957 (0.903 - 1.015)</td>
</tr>
<tr>
<td>Selected neonatal infections</td>
<td>1.008 (0.962 – 1.056)</td>
<td>0.74</td>
<td>0.977 (0.93 - 1.025)</td>
</tr>
<tr>
<td>Neonatal readmissions</td>
<td>1.025 (0.997 – 1.053)</td>
<td>0.08</td>
<td>1.002 (0.975 - 1.03)</td>
</tr>
</tbody>
</table>

Notes: 1. Trusts with > 1000 births per year
2. Sites with > 100 births per year
3. OR denotes odds ratio of busy periods (quintile 5) v. other times (quintiles 1 – 4) as reference; CI denotes Confidence interval
4. Accounting for trust/site effects using multi-level model

10.4.2.2. Consultant presence

Within the maternity extract, 51 of the 128 units (39.8%) with a recommended level of consultant presence were compliant. There were statistically significant differences in rates of perineal tears between compliant (2.95%) and non-compliant (3.29%) sites (Table 61). There were no statistically significant differences across the other measures and, likewise, further analysis against the revised
3-day neonatal readmissions ($p = 0.83$) and 3-day maternal readmissions ($p = 0.89$) showed no significant effects.

**Table 61. Association between compliance with consultant staffing levels and indicators of quality and safety of care**

<table>
<thead>
<tr>
<th>Measure</th>
<th>Unadjusted, % (number)</th>
<th>Adjusted OR (95% CI) (non-compliant vs compliant)</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Perineal tear</td>
<td>2.95 (3,916)</td>
<td>3.29 (9,443)</td>
<td>1.205 (1.001 – 1.450)</td>
</tr>
<tr>
<td>Puerperal infection</td>
<td>0.76 (1,345)</td>
<td>1.07 (4,149)</td>
<td>0.913 (0.535 – 1.557)</td>
</tr>
<tr>
<td>30-day maternal readmissions</td>
<td>1.10 (1,779)</td>
<td>0.95 (3,336)</td>
<td>1.006 (0.770 – 1.315)</td>
</tr>
<tr>
<td>In-hospital perinatal mortality</td>
<td>0.60 (1,026)</td>
<td>0.75 (3,076)</td>
<td>1.209 (0.852 – 1.716)</td>
</tr>
<tr>
<td>Injury to neonate</td>
<td>1.59 (2,513)</td>
<td>1.46 (5,481)</td>
<td>0.829 (0.460 – 1.494)</td>
</tr>
<tr>
<td>Selected neonatal infections</td>
<td>2.04 (3,468)</td>
<td>2.28 (9,342)</td>
<td>1.311 (0.789 – 2.177)</td>
</tr>
<tr>
<td>28-day neonatal readmissions</td>
<td>6.48 (10,022)</td>
<td>5.76 (21,400)</td>
<td>0.936 (0.730 – 1.200)</td>
</tr>
</tbody>
</table>

Notes: 1. OR denotes odds ratio of compliant v non-compliant performance; CI denotes confidence interval.

**10.5. Discussion**

**10.5.1. Summary of findings on seasonality**

We examined over 2 million deliveries across a three-year period. The longitudinal analysis showed that administrative data could be used to identify potentially important trends in obstetric care quality and safety. The analysis suggests that there are associations between month of delivery and outcomes. Notably, there are raised rates of neonatal readmission during the winter months (October to January). Interestingly, there is a substantial increase in early caesarean section rates in December, with one possible hypothesis being that NICE guidelines are not followed due to a lack of available slots to conduct elective caesareans during holiday season. There is no consistently poorer
month but I would suggest aiming for September and March would put the odds in your favour. The analysis provided some evidence to suggest there could be some poorer outcomes in July and August, a hypothesis that was rejected by a previous study on small preterm infants in USA.\textsuperscript{507}

10.5.2. Day of week

10.5.2.1. Summary of findings

The study demonstrated that babies born at the weekend have an increased risk being still born or dying in hospital within the first 7 days. Moreover, the study also suggests that other increased complication rates at weekends, with higher rates of puerperal infection, injury to neonate and 3-day neonatal emergency readmissions.

10.5.2.2. Impact: magnitude of weekend effect

I calculated, as described in the methods section, that the study suggests there are some 750 potentially avoidable perinatal deaths each year and an excess of 450 maternal infections if performance seen at the reference day, Tuesday, was replicated on other days.

There are also reasons to suggest that the inequality of care is more pronounced than identified in this study. If the effect is indeed caused by a staff deficiency and a lack of resources, then you would expect poorer quality and safety at all out-of-hours periods during the week, including bank holidays and weekday evenings and nights. If this is the case, then the out-of-hours periods during weekdays are masking some of the effect.

10.5.2.3. Amenable: resource levels

There are several possible explanations for these findings, including a lack of consultant obstetrician presence. This study provided some evidence to support the theory that one of the contributing factors to the ‘weekend effect’ might be failure to meet recommended levels of consultant presence. The association between consultant availability and performance also provides some evidence that the ‘weekend effect’ is amenable to the provision of healthcare.

On the analysis of the effect of activity on performance, there were not statistically significant variations, although the raised mortality at busy period might warrant further investigation. While based on a similar approach in the existing literature on measuring the association between levels of activity and outcomes, which concluded that hospitals that operate at or over capacity may experience heightened rates of patient safety events,\textsuperscript{508} a more sophisticated indicator for relative activity could be developed further.
10.5.2.4. **Precision**

The indicators identified a wide range of performance, by day of admission/birth, with statistically significant differences in four of the seven indicators, with further differences identified by disaggregating to day of the week.

10.5.2.5. **Minimum bias**

HES data do not include information on the time of admission and therefore I was not able to investigate the wider issue of the quality of out-of-hours care in this study. Other limitations should also be noted. Firstly, the administrative database used gives only limited information on the complexity of the delivery. Whilst I used a number of variables, such as Charlson index and number of previous admissions, to mitigate for any bias in case mix between admissions on different days of the week, there is no information on, for example, maternal obesity or smoking (as discussed in paragraph 9.6.4.1, p217). This may be important since a Canadian study identifying higher rates of early neonatal deaths from 1985-99 in weekends (OR 1.11, 95% CI 1.07 – 1.16) found these differences were no longer significant after risk-adjustment (0.96, 0.91 – 1.01), albeit another study found that adjusting for maternal characteristics had no material effect. A further limitation is that the administrative data can contain errors; however, you would not expect this to be different in the coding of weekday or weekend admissions.

I stratified one of the key results – in-hospital mortality – by mode of delivery to further investigate potential bias. For the most common delivery method, spontaneous vertex, the likelihood of perinatal mortality was higher at weekends than for weekdays, although this was not significant at the 95% confidence level (odds ratio: 1.059, 95% confidence interval: 0.983 – 1.140).

10.5.2.6. **Construct validity**

That many of the indicators identified evidence of an association between performance and day of delivery, and further analyses revealing lower level of specialist staffing also being associated a poorer outcome, provides evidence of construct validity. However, in lieu of having multiple comparable units – as with the cross-sectional analysis in the previous chapter – I have provided some assurance on validity by comparing consistency with previous literature. The results from this study are consistent with some previous results, including:

- Elsewhere in the UK, where a study in Scotland found adjusted weekend neonatal death of 1.3 (1.0 – 1.6, compared to weekday in-hours) which is similar for all out-of-hours 1.3 (1.1 – 1.6). Another study from the same country, while also finding the same effect, identified that a primary reason for the out of hours deaths was intrapartum anoxia.
Further afield, in Australia early neonatal deaths were much higher at weekends (by 29%, \( p < 0.001 \))\(^{504} \), and with a similar affect size (27%) found for neonatal deaths for infants born on Sundays.\(^{509} \) However, other studies, including those from Canada and USA rejected the hypothesis of greater complications at weekends.\(^{507,509} \)

### 10.5.2.7. Further work

Further work is needed to understand what organisational factors might influence the ‘weekend effect’ and to investigate centres that have reduced the disparities in access and outcome in out-of-hours care. A starting point for this is to allow hospitals to compare the extent of the ‘weekend effect’ in their organisation to that in their peers. Other future work to add to this analysis could include:

- Further sensitivity analysis by removing difficult cases and therefore reducing potential bias in case-mix. For instance, the analysis could be repeated excluding: cases with gestation age either missing or outside 37-43 weeks; and perinatal deaths ascribed to congenital abnormality or rhesus isoimmunisation, stillbirths and non-cephalic deliveries.\(^{510} \)

- Analysing the time of onset of the complication as well as the time of delivery. For instance, previous work has looked at the date of death as well as the date of birth\(^ {505} \) but there is difficulty in interpreting date of death from antepartum and intrapartum still births.\(^ {509} \)

- Looking at all out-of-hours periods, with inconsistencies in the existing literature on how to account for holiday periods.\(^ {445,503,505} \)

### 10.5.3. Conclusion

There is also scope to extend this analysis to other specialties, with similar results having also been found in a limited number of other clinical areas, such as pulmonary embolism, hip fractures and upper gastro-intestinal haemorrhage.\(^ {354} \) A greater understanding of the issue will also require better data, and the inclusion of an out-of-hours admission flag for hospital administrative data should be considered.

This chapter shows that it is possible to address some of the potential bias from case-mix and coding practice using longitudinal analyses; however, given the changes in coding practice over time which coincides with potential changes in service provision (and therefore hard to detect) this application is still susceptible to bias. The analysis by day of the week addresses most of the coding and risk-adjustment issues raised in the previous chapters and, therefore, permits more robust conclusions in what is, coincidentally, as a very high impact area.
Part IV:
Discussion,
recommendations
and appendices
Chapter 11.
Discussion

<table>
<thead>
<tr>
<th>Current measurement</th>
<th>Case for specialty-specific use</th>
<th>Review of current indicators</th>
<th>Application: validating common uses</th>
<th>Application: temporal analyses</th>
<th>Discussion</th>
</tr>
</thead>
<tbody>
<tr>
<td>Obstetrics</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Obstetrics</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Obstetrics</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
11.1. Summary

‘While most people are in favour of measurement, few are comfortable being measured.’ (Loeb 2004, p.6)

11.1.1. Overview

11.1.1.1. Quality measurement revisited

Using administrative data to measure the quality and safety of hospital care offers many opportunities. However, progress has been limited to few countries and predominantly to a small subset of broad measures, such as Hospital Standardised Mortality Rates. In this thesis, I investigate the potential advantages and feasibility – in terms of validity and applicability – of specialty-specific indicators.

Measurement of hospital outcomes is longstanding; at least dating back to the standardised mortality data published by Florence Nightingale in the early 1860s. Yet there remain calls for further development of indicators. The project described here addressed this, in part, through developing two sets of specialty-specific indicators (Chapters 4 and 8). As exemplified in the conclusions to a report on the failed Mid-Staffordshire NHS Foundation Trust, a greater consensus on how they are applied is needed. To address this, I outlined the key implications regarding the current use of indicators and summarised important methodological considerations in Chapters 1 to 3 and, subsequently, I tested and explicitly outlined key assumptions in applying the chosen indicators (Chapters 5 and 9).

11.1.1.2. Structure of this chapter

Given the shortcomings in quality measurement, coupled with the inexorable rise in the use of administrative data, it is important that researchers, practitioners and policy-makers focus on making best use of such datasets and related studies. The implications of the studies described here for researchers (see section 11.3), the collection of the data (11.4), clinicians (11.5) and policy-makers (11.6), along with some more general commentary from existing literature, are all covered in this discussion. These implications are preceded by a summary of the results in relation to the study hypotheses (11.1.2) and an overview of the methodological implications (11.2).

11.1.2. Summary of findings, by hypothesis

11.1.2.1. Hypotheses revisited

The project was designed to test three hypotheses (paragraph 2.7.1, p.59) regarding measuring the quality and safety of hospital care using specialty-specific indicators based on routinely collected
administrative data. Two of these hypotheses focused on the advantages of a specialty-specific approach (to create a bundle of indicators and better validate indicators) with the third, more generally, on the value of the temporal nature of administrative data. These propositions (in bold), along with a summary on how they were covered in the studies, are given below.

11.1.2.2. Hypothesis 1: Indicator bundles

A specialty-specific approach to developing indicators can result in a more useful and valid bundle of indicators to identify quality and safety issues. There has been a long-standing recognition that most efforts to measure performance have concentrated on mortality rates\(^ {50}\) and, consequently, there is a need to develop – for the purpose of monitoring hospital performance – additional non-mortality outcomes or process measures.\(^ {49}\) Four particular studies in the thesis supported this hypothesis, with the literature reviews (Chapters 4 and 8) showing high level of sensitivity for identifying indicators, while the experimental studies applying a sample of the indicators (Chapters 5 and 9) highlighted benefits in terms of additional validation opportunities.

The numbers of papers identified in the two literature reviews were also similar to those identified from an exercise to identify the broader category of all patient safety indicators (some 124 papers),\(^ {67}\) giving support to the assertion that the specialty-specific approach to identifying literature has good sensitivity (ability to identify relevant research). This is important since some of the prominent quality and safety initiatives, such as AHRQs patient safety indicators, have been developed using general (non-specialty) literature reviews and, as such, have not been based on a full assessment of existing evidence on, say, validity of existing measures.

It is briefly worth noting that as a rudimentary check, it was reassuring that the magnitude of abstracts identified through the search of electronic databases (1,999 for stroke; 1,670 for obstetrics) and full texts extracted (99 for stroke or 5.0% of abstracts identified; 80 for obstetrics or 4.8% of abstracts) were similar. This provides some evidence to suggest that the sensitivity and specificity of the specialty-specific literature search reviews were reasonable.

This ability to identify a more comprehensive range of indicators meant that I was able to apply broad indicator sets. I identified 20 categories of stroke indicators and 44 for obstetrics, and I applied 6 and 8 indicators, respectively, in the experimental studies. The advantage of applying bundles of indicators means that this project had higher content validity (the extent to which the measurement represents all facets of the area being investigated) than many existing studies identified through the literature reviews, which often applied just a single indicator, such as mortality rates, capturing only a very small proportion of the total possible complications. As health
services continue to improve, the focus on quality and safety is likely to continue to broaden – so moving further away from simple mortality – which lends support for the more holistic approach to measuring performance that I present here.

11.1.2.3. Hypothesis 2: Validation

Hospital-level comparisons could be improved through more robust validation techniques (coding practice, indicator comparison and organisational factors). Through the work, I demonstrated that investigating correlations between indicators, association with organisational factors, and the effect of coding practice, can all provide further assurance on levels of bias, construct validity, and amenability of quality and safety measures. The following results, in particular, are noteworthy:

- I was able to demonstrate that coding depth explained only some of the variation in performance, even across measures that are dependent on the recording of secondary diagnoses (i.e. aspiration pneumonia within stroke patients, injury to neonate and perineal tear) and, therefore, most susceptible to such bias.

- As well as providing more assurance on content validity (as discussed above), the study design – applying a bundle of indicators – had the advantage of allowing the evaluation of construct validity by comparing hospitals’ performance across all the measures. The vast majority of existing literature has failed to take advantage of this validation method, even when more than one indicator is applied. I was able to identify interesting associations; for example, trusts with higher scanning rates have lower stroke mortality (p < 0.05) and, for obstetrics, sites with higher maternal readmission rates also had higher levels of perinatal mortality (p < 0.001).

- There was some evidence to suggest that staffing is important to quality and safety. In particular, I found that lower midwifery staffing levels (i.e. higher birth-to-midwife ratios) were associated with higher perinatal mortality, and lower consultant presence within delivery wards associated with higher rates of perineal tears (both at p = 0.05).

While I showed that other applications (i.e. not benchmarking hospitals) hold advantages, as discussed in the following paragraphs, the investigatory work on validity of hospital-level comparisons is important since this remains a dominant use of these indicators.

11.1.2.4. Hypothesis 3: Temporal nature

The temporal nature of administrative data can be used to overcome some of the limitations of such indicators and highlight important quality and safety issues. Three studies presented here
(Chapters 6, 7, and 10) demonstrated how significant issues in healthcare can be identified using the longitudinal nature of the dataset in a robust way. This application of the indicators also has the advantage of controlling for the variations in coding practice over time and between hospitals. The evaluation of the association between day of admission (or delivery, for obstetric measures) and complication rates was the first comprehensive assessments of this issue in England for these two specialties by applying multiple indicators so having higher content validity.

As well as controlling for some of the possible bias, the temporal application also suggested significant shortcomings in care. For example, 7-day stroke in-hospital mortality was 10.6% at weekends compared to 8.9% at weekdays (adjusted OR 1.178, 95% CI 1.120 – 1.240). Similarly, the perinatal mortality rate was 7.3 per 1,000 at weekends, some 0.9 per 1,000 higher than at weekdays (1.071, 1.019 – 1.126). By matching cases between out-of-hours and in-hours cases, I was also able to estimate the scale of avoidable harm. For instance, in stroke care, the study suggested there are, each year, some 250 potentially avoidable in-hospital deaths within 7 days and an additional 650 people could be discharged to their usual place of residence within 56 days if the performance seen at weekends was replicated at weekends.

I was also able to use the temporal nature of administrative data to evaluate the existence of seasonal trends in obstetric care and the effect of a regional reform of stroke services. In both cases, I found potentially important differences in the quality and safety of care with, for example: a significant reduction in 7-day stroke mortality within the region that had undergone a service restructuring, in comparison to other areas; and raised rates of neonatal readmissions during winter months.

11.2. Methodological issues

11.2.1. Introduction to study limitations

Methodological issues related to the specific studies were raised in the previous chapters’ discussions. However, some general limitations about the approach taken are worth noting when considering the implication of the work as a whole.

11.2.2. Limitations to development of indicator sets

11.2.2.1. Addressing limitations in review process

The limitations of this review process include the potential bias introduced from the use of a single reviewer. No formal meta-analysis could be undertaken because of the heterogeneity in study methodologies and underlying data frameworks and I did not use a template to assess the quality of
the papers since there was a very diverse range of purposes across the studies. Despite the limitations, the literature reviews shared the features of a systematic, rather than narrative, review (paragraph 3.4.2.3, p72). However, there could be benefits to future research in using a more resource-intensive review process which, at least, could improve face validity of results.

11.2.2.2. Selection of indicators

One limitation of the framework was in the bias towards measuring what has previously been measured. Whilst having peer-reviewed articles on particular indicators may support face validity, it may bias future work towards such indicators. There would be value in complementing such literature reviews with further exercises to develop indicators in areas which are not covered by the existing measures.

The design of the study was, in the first instance, based on a prospective analytical framework, and supplemented reactively – formed on the basis of the gaps in the literature. I used a range of experts, from coding specialists to medical consultants to help develop the indicator set; however, I did not use any formal approach to reach consensus. This could be considered in future research; for instance, if there are sufficient resources available an approach such as RAND’s Delphi method could be applied.

11.2.3. Indicator validation

11.2.3.1. Lack of ‘gold standard’

Ensuring that indicators are not just easy to apply, but also have appropriate validity, is important: “something that’s measurable may not be worth measuring, and maybe you can’t measure the things that are worth measuring. What damage do you do by releasing information just because you can measure it?” (GP quote, p 1280) A review, published in 1998, by Fitzpatrick and colleagues outlined the key aspects of validity of patient-based outcome measures albeit in the context of use in clinical trials. The approach used in this study tallies well against this and other such frameworks. However, many aspects and tests of validity (e.g. concurrent, convergent and discriminative) depend on the existence of a ‘gold standard’ to provide a basis for comparison; if no gold standard exists, they represent a form of construct validity in which the relationship to another measure is hypothesised. It is, therefore, an inherent weakness that no absolute, incontestable statement on validity of indicators can be made here.

11.2.3.2. Minimum bias

This study provided the most comprehensive assessment to date on the effect of coding practice. While the analysis on both specialties suggested that key issues around variations in coding practice
only partially explain variation in performance against the indicators. However this was not comprehensive, with other potential variations in coding practice remaining untested.

The application showed that you can adjust for the measured case-mix factors by building converging regression models. However, the extent to whether this is sufficient remains debated with some suggesting that good risk adjustment is possible whereas others, for instance, highlighted that there is bias in the coding of medical records which has implications on the administrative data derived from it. The lack of some key clinical factors within administrative data is an inherent limitation when comparing units with different case-mixes.

11.2.4. Limitations to findings on shortcomings in care

11.2.4.1. Reference day
As highlighted in some of the exploratory work on stroke care trends (paragraph 6.5.2.2, p146), the choice of reference period when comparing performance has a significant effect on the derived odds ratios. One inconsistency in my approach to the stroke and obstetric out-of-hours analysis should be noted. For the latter I used Tuesday (rather than Monday, as for the stroke analysis) as the reference day. This decision was made a priori as, on Mondays, large numbers of babies are born to women who have received care during the weekend and, as such, this day does not represent a typical in-hours day. Ideally, the analysis would be able to disaggregate between all in-hours and out-of-hours periods but the data used does not include a time field (rather only date) and this is covered later (11.4.2.6, p260).

11.3. Future research

11.3.1.1. Introduction to future research
Some specific areas of future research are covered in early chapter discussions. A few notable areas are outlined here, covering: the development of indicator sets within other specialties; areas for further research within stroke and obstetric specialties; the development of reporting standards for similar research; and an outline of which study designs for such indicators might be fruitful.

11.3.2. Developing indicator sets for other specialties

11.3.2.1. Introduction
An obvious area for expanding this project is to transfer the approach for identifying indicators to other specialties. This section covers why the specialty-specific approach was worthwhile, then sets out a framework for considering what specialties to choose for similar interrogation, and finishes by setting out the importance of publishing the resultant indicator set.
11.3.2.2. **Advantage of the specialty-specific approach**

The specialty-specific reviews proved useful to developing an indicator set. For both specialties, the increase in research has been similar (50% per annum for stroke, 60% for obstetrics) with few studies prior to 1990 (as expected since few large administrative datasets were available before then, as highlighted in paragraph 2.1.1.2, p39. This increase is likely to reflect both the growth in agreed indicator sets, which even if not directly relevant to the specialties supports a culture of using administrative data to monitor performance, and increasing access to these databases. Both specialties yielded a substantial number of indicators (as discussed below), with many of these appearing regularly in a range of the literature, including peer-reviewed articles. The implication is that future research on developing indicator sets to measure performance should not rely solely on generic searches of indicators and instead consider using specialty-specific terms.

Similarly, I found there are advantages of looking at more than one specialty. For example, I was able to learn from different maturation of each specialty’s research, and applied a technique for evaluating coding practice that had been used in one of the specialties (stroke) to the other specialty (obstetrics).[^263] I focused on two specialties, both within acute care, and while this scope was shown to be appropriate – with proven importance and difference between specialties – not all lessons will be generalisable. The implication is that there is value in expanding the application of indicators to other specialties.

11.3.2.3. **Choosing other specialties**

Making the assertion that there is value in expanding this approach to other specialties also raises the question of which would be the most fertile area to expand into. Table 62 shows that comparing the two specialties does not give a clear direction on whether one specialty outperformed the other as a subject for such reviews. The high number of obstetrics indicators could suggest high content validity by representing more facets of the subject of evaluation. Yet for the obstetrics study, indicators are limited to perinatal care whereas the stroke analysis included indicators across the pathway. For the former, this suggests reasonable content validity if perinatal care is the subject but not necessarily the case if all obstetric care. While not giving a prioritisation for which specialties to apply this approach to next, this work does provide a framework to compare facets of specialties (Table 9, p69) and shows that relatively disparate specialties can both provide interesting insight though the literature-review process. That said, development should be directed to account for both clinical importance (e.g. impact and amenability) and methodological feasibility (e.g. scientific soundness and applicability).
Table 62. Systematic review findings in relation to analytical framework

<table>
<thead>
<tr>
<th>Domain</th>
<th>Both</th>
<th>Stroke</th>
<th>Obstetrics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Importance</td>
<td>(+) Clearly articulated policy importance and known issue</td>
<td>(+) Evidence of effect of organisational differences</td>
<td>(+) More benchmarks of acceptable performance</td>
</tr>
<tr>
<td>1. Impact</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. Amenable</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. Face and content validity</td>
<td>(+) Range of (peer-reviewed) indicators</td>
<td>(-) Fewer measures</td>
<td>(-) Indicators focus only on perinatal part of the pathway</td>
</tr>
<tr>
<td>4. Precision</td>
<td>(+) Known variation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5. Construct validity</td>
<td>(-) No “gold standard”</td>
<td>(+) More process indicators to supplement outcome measures</td>
<td>(+) More indicators focused on the same construct (i.e. more opportunity for convergent validation)</td>
</tr>
<tr>
<td></td>
<td>(-) Few bundles used</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6. Minimum bias</td>
<td>(+) Model fit often described</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(-) No consistency or transparency in risk model</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(+) Many indicators applied to administrative data</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7. Data availability</td>
<td>(+) Examples of administrative data being applied</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(-) Few studies on ICD-10 data</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8. Reporting burden</td>
<td>(+) Many examples of previous reporting</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9. External and ecological validity</td>
<td>(-) Predominantly based on USA data and ICD-9 coding framework</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(-) Lack of transparency</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Notes: The positives (+) for one of the specialties can be interpreted as negatives (-) for the other specialty in this comparison. These converse arguments are not presented for brevity.

11.3.2.4. Publishing indicator sets

The literature reviews also provide some evidence to suggest that areas with published indicator sets have benefited accordingly from increased volume of, consistency across, and evaluation within, research. Whilst comparisons between the numbers should be treated with caution, a crude cross-check of the categories of indicators identifies more within obstetrics (42) than in stroke (20). The difference may be merely the natural result of the coding frameworks being more appropriate for identifying obstetrics complications and the fact that for this specialty there are two sets of data (the mother and baby records). However, this could also reflect the influence of the widely recognised indicator sets, with a number of obstetrics indicators appearing, for instance, the AHRQ PSIs. This latter hypothesis is supported by the fact that there was more regularity in the use of indicators in these sets even outside their country of creation. If true, the implication is that developing and publishing indicator sets can create a culture of measurement so proliferating indicator numbers and
improve the use of administrative data, increasing the amount and consistency of research and, resultantly, allowing better benchmarking of performance (indeed, I found that there were more published indicative benchmarks for acceptable performance against the obstetrics indicators than for the stroke research).

11.3.3. Further research within these specialties

11.3.3.1. Introduction to further research within stroke and obstetrics care

The stroke and obstetrics studies highlighted some areas for further research within these specialties. In particular, within the two specialties, there would be value in: conducting further work to improve the indicators; applying the indicator set used here in other countries; and better understanding some of the associations between organisational factors and performance.

11.3.3.2. Improving the stroke and obstetrics indicators

As stated above, the studies in this thesis suggested need for further research on validity, which is echoed in the existing literature. The utilisation of such measures – whether, for example, for internal benchmarking by hospitals or for regulation – must be in proportion to the confidence over the validity of the individual indicators. So, as well as tempering current reporting to the present validity of indicators, efforts should also be made to improve the validity of the indicators. Indeed, in addition to developing new indicators (as per section 11.3.2), there have been calls to further develop those that already exist. In fact, the WHO stated that “existing measures should be refined and further validated.”(p11)515

In the editorial introduction to the first major set of quality indicators – CSP, Iezzoni et al (1992) – the commentator states that they “believe the algorithms will prove difficult to improve upon.”(p361)50 History has proved this statement to be wrong with incremental improvements in the approaches to using administrative data (as well as some retrograde steps) having been made, and this monograph continues the important development. In terms of validating the indicators applied to the data, again the existing literature across the two specialties was limited with the most common techniques using different risk-adjustment models and comparing results to other sources. Further work should continue across all the aspects of validity highlighted in this study.

11.3.3.3. Application of the two indicator sets to other healthcare services

For both specialties, the majority of studies were based on data from the USA and using ICD-9 data, which is consistent with previous literature on measures derived from routinely-collected datasets.137 The geographical coverage of the obstetrics’ studies was greater than in for the stroke research; however, this may reflect differences in the search strategies, with the obstetrics review
able to include foreign language articles. Some of the indicators and findings presented here might be specific to the NHS and its datasets and so further research should focus on less well-researched health services.

11.3.3.4. Addressing specific research gaps
Across stroke and obstetrics research, many studies have focused on the association between performance against the indicators and different hospital characteristics. However, while stroke indicators were used to evaluate different service types, such as the use of stroke units, no studies based on administrative data had looked explicitly at different modalities for obstetrics delivery. This may reflect homogeneity in the organisation of services in some countries, although this explanation does not hold everywhere with, for instance, the proliferation in midwifery units – as an alternative to hospital obstetrics units – in England. The challenge for such research – as highlighted in particular on the discussion on evaluating stroke services in London and effect of size of stroke units – is to fully recognise that different modalities might have heterogeneous case-mixes and not all such factors might be measurable within the data.

More generally, the research presented in this thesis found some evidence to suggest performance is associated with staffing levels. Further work is required to confirm these relationships, including an exploration of the cause.

11.3.4. Developing reporting standards

11.3.4.1. Transparency as a methodological strength of this study
Citing possible bias due to differences in hospital-level coding has become almost a default, generic limitation to include when writing up research on the use of administrative data to compare between providers performance. However, this project aimed to quantify this potential bias. More widely, this study benefited from being explicit about assumptions, methods and model fit.

In comparison, across the existing literature, there was a similar lack of consistency and transparency in the strategies used for case ascertainment for both the obstetrics and stroke research. Likewise there were a plethora of strategies used for accounting of potential patient confounders, ranging from risk-adjustment models including only age to others combining diagnosis, procedure, demographic and case-mix factors.

11.3.4.2. Recommendation to develop reporting standards
Given the lack of consistency in the application and reporting of studies using administrative data, a natural conclusion would be to introduce a checklist or similar process to standardise the research
and, therefore, improve interpretability and generalisability. Indeed, in 2002 a study by the Agency for Healthcare Research and Quality called for more work on identifying and resolving quality rating issues for observational studies. Generic existing scoring system on quality of studies – such as Grading of Recommendations, Assessment, Development and Evaluations (GRADE) – lack the specific requirements needed to add transparency to studies using administrative data for which non-randomness places an importance on risk-adjustment techniques. In fact, even checklists on observational studies such as Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) guidelines fail in this same respect. Such research might start with the most recent version of the Cochrane Handbook which contains guidance on dealing with non-randomised studies in systematic reviews of interventions, from the protocol to synthesis stages. In particular, such reporting should include details on: data fields interrogated; diagnosis and procedure codes used in the indicator algorithms; the case-mix model fit; and validation techniques applied.

11.3.5. Study designs

11.3.5.1. Introduction

I applied the indicators using a range of different experimental designs to evaluate the specialty-specific indicators based on routinely-collected hospital administrative data. In all these applications, and across both specialties, the studies revealed that the area was important (e.g. through estimating potentially avoidable deaths) and that performance could be discriminated (e.g. the high numbers of statistical outliers); however, each design has inherent the strengths and weaknesses as summarised below (Table 63, with a more comprehensive assessment of performance of each approach against the analytical framework given in the respective chapters). A few key study designs which could be fruitful areas for further research are described in more detail below.
Table 63. Summary of strengths and weaknesses of the hospital and patient-level approaches

<table>
<thead>
<tr>
<th>Study design</th>
<th>Hospital-level analysis</th>
<th>Longitudinal</th>
<th>Regional reform</th>
<th>Day of week</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Hospital-level analysis</strong></td>
<td>Cross-sectional</td>
<td>Cross-sectional</td>
<td>Quasi-experimental</td>
<td>Cohort</td>
</tr>
<tr>
<td><strong>Key strengths</strong></td>
<td>Discriminates even after accounting for some coding variations</td>
<td>Hospital-level trends will control for some hospital level influences on coding</td>
<td>Ability to compare with both historical performance and that of control group</td>
<td>Unlikely to have variation in coding by day of the week</td>
</tr>
<tr>
<td><strong>Key limitations</strong></td>
<td>a. Bias from unaccounted variations in coding by hospitals</td>
<td>a. Evidence that coding practice is changing over time</td>
<td>a. Case-mix changes due to reorganisation</td>
<td>a. Possibility of unmeasured variation in case-mix by day of the week</td>
</tr>
<tr>
<td></td>
<td>b. Bias within regional networks with added heterogeneity of patients between hospitals</td>
<td>b. Variation in case-mix over time</td>
<td>b. Uncertainty of counter-factual</td>
<td></td>
</tr>
</tbody>
</table>

**11.3.5.2. Epidemiological studies**

There is a longstanding understanding that the patient factors (e.g. co-morbidity and age) and hospital factors (e.g. coding and numbers of caesareans performed) associated with adverse events are important. Yet epidemiological studies are difficult, with challenges to interpret findings when there are biological explanations for risk factors affecting outcomes and, specifically on using administrative data, there are weaknesses with, for instance, poor recording of some key risk factors.

A literature review conducted in 2003 concluded that analysis of administrative data was not relevant to the purpose of understanding the latent causes and contributory factors behind adverse events. However, since this review a number of patient factors have been analysed using administrative data, including significantly higher rates for some adverse events for: children, schizophrenic patients, certain races and ethnicities, and females. Similarly, significant associations with hospital factors have been identified, including lower rates of some adverse events.
Part IV: Discussion, recommendation and appendices

for providers with: more sophisticated and mature IT infrastructures performed significantly better on the largest number of PSIs;\textsuperscript{123} access to critical care;\textsuperscript{124} and holding daily ward rounds.\textsuperscript{518} For better understanding of the risks and to permit future benchmarking, further work should focus on gaining better understanding of hospital factors.\textsuperscript{519}

11.3.5.3. Quasi-experimental

Such indicators offer the potential for undertaking trend analysis.\textsuperscript{62} As a result, it is possible to assess whether an initiative, such as implementing an IT solution,\textsuperscript{122,123} has had an effect on patient safety. And there are examples which show that rates of adverse events can be influenced by initiatives; one project found that improved staffing reduced errors in a 24 hour paediatric critical care satellite pharmacy with unit dose drug distribution.\textsuperscript{14} However, such positive findings might be the result of the Hawthorne (observer) effect whereby people modify their behaviour when new initiatives are introduced. A weakness of such quasi-experimental approach is that coding changes over time, thus making disaggregating underlying effects more difficult.

Despite limitations in such study designs, researchers should take advantage of the longitudinal and comprehensive data which allow for time- and cohort-based comparisons. This thesis has already presented examples of evaluating interventions, such as those existing studies on implementing stroke unit care and the original work undertaken on investigating London’s regional reorganisation. However, there are many other techniques – such as synthetic control matching – and numerous other possible foci for further studies.

11.4. Data implications

11.4.1. Background

11.4.1.1. Study implications on data

While the studies here demonstrated that key coding issues could only explain some of the variation in performance, it did highlight the extent of variation in coding practice. For example, providers’ average number of distinct diagnoses codes per admission ranged from 5.0 to 10.7 within stroke care (by hospital trust) and from 2.8 to 6.6 for maternal admissions (by hospital site). Similarly, the use of the ‘stroke unspecified’ diagnosis code ranged from 0.2 to 42.6% of admissions across hospital trusts. Some of this variation may be explained by variations in patient case-mix but the huge differences suggest that there are also differences in coding practice.
11.4.2. Framework for improving data

There is a need to gain understanding of why data errors are created and, to this end, Benin and colleagues have provided a useful framework by further disaggregating the potential errors as: categorising error (non-recordable differences in case-mix); entry error (improper inputting of the data); and query-error (misunderstanding of the data schema). Further work is required to ensure coding within HES is sufficiently accurate, including improved and more accessible coding guidance, and mandatory recording of specific procedures and diagnoses. Where discrepancies are found in the recording of data, there is a need for further research for both quality and safety auditors and regulators (to ensure performance is known) and hospitals (given the implications on their remuneration regarding, for instance, the coding of thrombolysis and scanning).

11.4.2. Recommended interventions

11.4.2.1. Peer review of coding

To improve consistency in application of the coding rules, there should be a degree of peer review on practice. Given the changes in the governance arrangements around the quality of data given the forthcoming abolishment of the Audit Commission (who previously had a role in auditing this data) such a peer review process could be incorporated into any new data validation process.

11.4.2.2. Adjust the coding framework

The Government has proposed that the coding framework for administrative data may be updated. This study suggests that, if administrative data are going to be used for monitoring healthcare performance, consideration should be given to ensuring the framework allows for the recording of a wider range of quality and safety issues.

11.4.2.3. Reassess coding rules

There should be a review of whether some of the coding rules, such as the guidance to not record procedures undertaken before the decision to admit has been taken, should be amended. In particular, this current rule could introduce some bias where hospitals have differing procedures for admitting patients and might also result in an underestimate in, for example, scanning rates.

11.4.2.4. Present on admission

A commonly cited issue with using administrative data, is attribution with, for instance, potential errors when adverse events are present on admission. One solution is to enrich the administrative data, such as adjusting for conditions present on admission can be useful. The identification of diagnoses that are present on admission can be addressed through linking to other datasets, such as laboratory results. This endeavour of linking datasets also has the advantage of potentially allowing
Part IV: Discussion, recommendation and appendices

a greater number and detail of indicators, so improving many aspects of validity. Alternatively, a present-on-admission flag could be introduced as used in some systems in the USA and Australia.

11.4.2.5. Do not resuscitate
As with ‘present on admission’, there are potential advantages in being able to flag up, within the data, patients whose status is Do Not Resuscitate (DNR).\textsuperscript{520} For instance, a previous publication suggested that removing DNR cases from the denominator might improve indicator performance.\textsuperscript{149} However, this particular amendment to the coding framework should be done with consideration to both the fact that a DNR flag could reflect an adverse outcomes from poor care (and so such cases shouldn’t be excluded from quality measures) and the risk of incentivising gaming.

11.4.2.6. Out-of-hours identifier
The research I conducted in both specialties produced findings that were consistent with poorer standard of care at weekends. However, the true extent of inequality in outcomes by time of admission might be masked by a lack of time-, rather than just day-, stamp. Such a flag should be considered since it would allow researchers to identify the effect of different access to physical resources and staff during out-of-hours periods.

11.5. Clinical implications

11.5.1. Clinical guidelines and quality improvement
For most of the measures applied in this thesis, there is no clinical consensus or guidelines for what are acceptable levels. In some instances this may be appropriate but in others this information is potentially useful to clinical teams and services because it supports the monitoring of performance. The two exceptions (where measures had consensus/guidelines) – with both coincidentally having deficiencies in performance – are:

- access to a scan for people who have had a stroke, for which extant guidelines recommend that all such patients should receive brain imaging “within a maximum of 24 hours after onset of symptoms”.\textsuperscript{334} The actual performance is that 69.7% of patients receiving a scan within one day of admission.

- on caesarean sections prior to 39 weeks, for which the extant guidelines state that “the risk of respiratory morbidity is increased in babies born by CS before labour, but this risk decreases significantly after 39 weeks [and] therefore planned CS should not routinely be carried out before 39 weeks”(p18).\textsuperscript{490} The current rate is around a third of elective caesarean sections being carried out before 39 weeks. While some of this apparent
shortcoming may be due to poor coding, the seasonal differences (with particularly high rates of elective caesarean sections during December) suggest there are clinical shortcomings.

These two cases, with large proportions of people being treated in contravention to the clinical guidelines, seem unacceptable. The implication is that the clinical guidelines need to be revisited, checked to ensure they are deliverable and appropriate, then re-communicated.

### 11.5.2. Advantages of reporting to a ‘specialty team’

#### 11.5.2.1. Introduction

The studies clearly identified findings of clinical importance and, therefore, the implication is that clinicians need to have these findings reported to them.

Health services should consider how these specialty-specific indicators could be reported routinely and directly to the specialty team. This reporting style would represent a hybrid of the administrative data identification method and ‘morbidity and mortality conference reporting’ format (where the clinical team formally discuss adverse events) and, in doing so, could harness the benefits of both. The motivation for reporting to clinicians is implicit in the arguments set out for reporting to specialties directly (section 2.6, p54). Further to these, the literature does set out further motivation for feeding back the results from indicators directly to specialty teams, as summarised below. Most refer to clinicians but hospital managers or even coders in that specialty should not be overlooked for risk of entrenching any interdisciplinary divide.

#### 11.5.2.2. Learning and culture

Huber’s review of organisational learning research makes a distinction between experiential learning and “learning from searching and noticing”. The former type of learning is the category under which clinicians learn from their mistakes, whilst the second category would include the use of indicators to help clinicians prospectively improve through understanding the risk factors associated with patient safety. Certainly, there is recognition that reducing adverse events requires in-depth involvement of clinicians. Leggat and colleagues (2008), for instance, point out the importance of the skill and capacity of healthcare workers in delivering safer patient care. Similarly, one of the keys to improving patient care is to ensure that the health workforce develops the will to address the issues, and clinicians are well placed to change cultures.
11.5.2.3. Develop new and existing indicators

The specialty team members are in a good position to propose indicators. The increased demand for indicators that has precipitated from the raised profile of quality and patient safety could be addressed, at least in part, by developing and validating more indicators.\textsuperscript{73,523} Further to meeting the need for additional information, developing additional indicators would also:

- reduce the focus on particular diagnoses and whether they are coded. This would diminish any ability and incentive to ‘game’ (as discussed in §11.6.1.2) by not coding accurately or comprehensively;

- would create the opportunity to test correlations and identify factors from among indicators. This in turn would enable the indicators to depict quality and safety on a number of distinct dimensions;\textsuperscript{73} and

- reduce reliance on existing research, which I highlighted earlier as a methodological issue.

The specialty team can also focus new indicators on areas where they believe there is scope for improvement and so meet calls that measurement tools should focus on whether any harm caused was preventable.\textsuperscript{71}

11.6. Policy implications

11.6.1. Variations in performance

11.6.1.1. Scale of variation

There were large variations between providers in their performance across the indicators. For example, same-day scanning rates varied from 20.4\% to 79.3\% between hospital trusts and early caesarean section rates varied from 1.1\% to 60.7\%. Even after accounting for random chance, there was a strong indication of substantial underlying differences. There were 181 occurrences of hospital trusts performing statistically significantly differently to the average at the 99.8\% confidence level across the 6 stroke measures. For the 8 obstetrics indicators there were 441 such outliers. Moreover, differences in coding practice appeared to only partially explain the variation in performance.

11.6.1.2. Use of benchmarking

These findings raise questions on how such research should be used. Benchmarking – or comparing performance against peers – can be done either confidentially or publicly. There are cases of specific indicators being used for public reporting of hospital performance, by benchmarking against similar
providers. In the NHS, a number of indicators of quality and safety, derived from administrative data, are now used to benchmark providers and commissioners of care and indeed the publication of results from this study builds on this. However, a number of problems associated with this possible utilisation have been pointed out, including that:

- if indicators are used to evaluate individual hospitals and in the absence of an effective system for auditing the recording of adverse events in the administrative data, there is a risk that hospitals may stop coding incidents that could implicate the occurrence of a patient safety event;\(^7\) and

- due to the limitations on validity, a report could show an unduly poor rate and, if not confidential, a hospital might be pressurised into making a knee-jerk reaction to correct this reported problem.\(^1\)

Given these unresolved issues, inter-institutional comparisons should be “approached with caution” (p788),\(^1\) although Marshall and Romano argue that the concept of publicly available data is becoming more acceptable to both clinicians and managers.\(^5\) A study based in the NHS, which involved sending out benchmarking reports to 170 hospital trusts, reported, however, that “trusts were generally positive about receiving the reports” (p13).\(^1\) The pressure to publish benchmarking data to support policies such as choice of provider, which exists in many health services, means that benchmarking is likely to remain a common application and, as such, I committed a large proportion of my research into investigating some of the issues around this application. As discussed earlier, this study suggests that some of the commonly cited limitations of benchmarking – such as coding depth – can be addressed to some extent.

### 11.6.1.3. Acceptability of variation

The providers of health services in England are increasing autonomous, although there remains a quality regulator (the Care Quality Commission, CQC) with responsibility for registering, monitoring and inspecting providers. Some level of variation may be inevitable but it is unclear what level policy makers deem acceptable. The scale of the variation is sufficient to suggest that providers and the CQC should consider how to address some of the more significant variations. One action could be to use such indicators to attempt to reduce such variations and this is covered in the next section.
Part IV: Discussion, recommendation and appendices

11.6.2. Addressing issue on association between complications and day of the week

11.6.2.1. Scale and causation
This project identified significant variations in out-of-hours performance, as outlined earlier in the discussion. The stroke analysis benefited from reporting both process and outcome measures for this weekend analysis which provides greater construct validity and also from certain process measures being less susceptible to bias from case-mix. As such, policy makers should ensure that further studies on out-of-hours care should endeavour to include process measures. Further research should also continue to look at this issue through a specialty-specific lens, with previous research suggesting a heterogeneous effect across specialties, which could result in significant shortcomings of care being masked by a broader scope.524

11.6.2.2. Addressing the ‘weekend effect’
In England, in response to findings on quality of care at weekends, a number of initiatives have been developed to reduce mortality, increase efficiency, ease access, and ensure patients receive the same standard of care regardless of the day of the week. Further research suggested that achieving a seven day service would: require local solutions; cost up to 2% of total patient care income; and unlikely to be cost-neutral.525 In a resource constrained health service, this equation has clear policy implications, requiring health service leaders to evaluate how much value they place on providing an equal level of care at all times.

11.6.3. Application of indicators by policy makers

11.6.3.1. Scope of discussion on application of indicators
The previous ten chapters have set out the rationale, methods and considerations relating to the application of specialty-specific indicators. Coupled with the proliferation of indicator use, it would therefore be remiss not to set out some key issues regarding the use of these indicators to improve quality and safety. The work represents only a narrative summary of the multitude of considerations and views expressed elsewhere by a range of disciplines, including that of Benn and colleagues which draw on psychology,126 but aims to give an overview of the considerations for policy-makers regarding how such indicators could be applied. The similar discussion for researchers – on experimental design (e.g. quasi-experimental or epidemiological) – was covered earlier.
11.6.3.2. Revisiting reporting

The application and interpretation of indicators based on administrative data depends on a multitude of often dependent factors, such as the purpose and type of event being identified (Figure 32).

**Figure 32: Potential uses of patient safety indicators**

<table>
<thead>
<tr>
<th>Adverse clinical outcomes</th>
<th>Side effect</th>
<th>Complication</th>
<th>Misadventure</th>
<th>Adverse event</th>
<th>'Violation'</th>
<th>Preventable adverse event/error</th>
<th>Sentinel event</th>
<th>Negligence</th>
<th>Malpractice</th>
<th>Malicious injury</th>
<th>Homicide</th>
</tr>
</thead>
<tbody>
<tr>
<td>Patient safety</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Public Reporting</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Compensation</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Professional regulation</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Criminal justice</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Source: Based on Australian Centre for Economic Research on Health (ACERH)

Of these broad categories of purposes, the following specific examples of applications have been suggested: document the quality of care; make comparisons and benchmarking over time between places (e.g. units, hospitals); make judgments and set priorities (e.g. choosing a hospital or surgery or organising medical care); support accountability, regulation, and accreditation; support quality improvement; and support patients’ choice of providers.65

The WHO has recommended that reporting systems should be blame-free, concluding that successful systems are: non-punitive; confidential; independent; expert analysis; credible; timely; systems-orientated; responsive.97 This conclusion was echoed by the team constructing a national set of indicators in Denmark which called for reporting systems to be: confidential, sanction-free, learning oriented.117 Both of these recommendations include ‘confidential’ but this may be at odds with policies to improve transparency, such as the Freedom of Information act in the UK, which highlights the conflict between theoretically-attractive and practical solutions.

11.6.3.3. Types of application

The potential application of indicators was covered briefly in the section 2.5 (p51-) to explain the context with which the indicators are used. I give greater detail, including some commonly cited
strengths and weaknesses (summarised in Table 64), of some of these applications below drawing predominantly on the literature on generic (non-specialty-specific) indicators. The limitations and strengths of benchmarking between providers were covered earlier (paragraph 11.6.1.2, p262) and are not repeated here.

As a WHO report identified, whilst “they [PSIs] have not yet been validated extensively,... they have been found useful in several studies for estimating rates, identifying risk factors and tracking trends in inpatient safety events.” More generally, a review of patient safety recommended that indicators “can serve as timely indicators for latent problems that the organization is not otherwise tracking” (p1644). My view would be to use more robust applications for benchmarking adverse events and perhaps even accountability, with other less-valid applications for case finding, with any use of such indicators within a wider portfolio of measures. I did not identify sufficient evidence within the literature that using such indicators for reimbursement or policy-making would necessarily be effective.

Table 64. Key strengths and weakness of common applications of administrative data

<table>
<thead>
<tr>
<th>Application</th>
<th>Strength</th>
<th>Weakness</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accountability</td>
<td>11.6.3.4 Indicators can be aggregated to whichever level of accountability is being evaluated</td>
<td>Majority of commentators warn against ‘hard’ application due to, for instance, gaming</td>
</tr>
<tr>
<td>Benchmarking</td>
<td>11.6.1.2 Administrative data cover many providers (or other units) and are longstanding (so can compare over time)</td>
<td>Evidence of variation in coding practice</td>
</tr>
<tr>
<td>Reimbursement</td>
<td>11.6.3.5 Easy to implement as data often forms part of funding framework</td>
<td>Strong incentive to bias coding</td>
</tr>
<tr>
<td>Priorities</td>
<td>11.6.3.6 Linked to activity and cost</td>
<td>Lack of comparability between specialties due to coding framework</td>
</tr>
<tr>
<td>Case finding</td>
<td>11.6.3.7 Promotes openness</td>
<td></td>
</tr>
</tbody>
</table>
11.6.3.4. Accountability

To assess how indicators could be utilised to hold people to account, one has to consider to whom the assessment will be made. An advantage of administrative data is that the indicators can be aggregated to different levels to meet these difference reporting needs. In England, the levels of aggregation have focused on organisations or physicians. Such transparency has, for example, been called for in reports on adverse care at Bristol Royal Infirmary and Mid-Staffordshire NHS Foundation Trust. However, as also suggested earlier, a review of the literature suggested that any patient safety measure should be non-punitive, so this application remains contentious.

If health systems start to employ indicators for collecting ‘hard’ evidence to apportion accountability and responsibility then there will be a perverse incentive to under-report adverse events and, therefore, reduce the validity of the underlying dataset for the indicator. There is a need for a reporting culture and, to catalyse a move to such a culture, there will need to be an end to any perceived or actual blame culture. This argument leads to the conclusion that indicators should not be used to hold people accountable. This is particularly important since many adverse events may be blameless. Certainly, “in aviation maintenance – a hands-on activity similar to medical practice in many respects – some 90% of quality lapses were judged as blameless” (Marx 1997 cited in 527).

Given the level of specificity and sensitivity of current indicators, they could only be applicable for compensation, professional regulation, or criminal justice if used in conjunction with other methods. Certainly, many academics have warned that current indicators should not be used as embarrassment, let alone blame or sanction.

11.6.3.5. Reimbursement

Another potential use for these indicators is to determine levels of reimbursement for healthcare providers. In fact, in the USA if the administrative data reveals that one of a specific list of sentinel events has occurred during an admission the hospital will not be reimbursed for the costs of the healthcare for the index event or for treating the harm relating to the adverse event. However, rate-based indicators have rarely been used to determine levels of reimbursements. The issues relating to benchmarking that were outlined above also hold for this use. Bahl and colleagues have concluded that indicators based on administrative data are not appropriate to profile hospital performance or determine levels of reimbursement. Indeed, Davies, in explaining the refinement of the HCUP quality indicators, recommended that indicators should not create incentives to improve measured performance without truly improving the quality of care provided, which could be an unintended consequence of linking with reimbursement.
11.6.3.6. Make judgements about priorities

Priority setting – deciding about the areas to invest resources and focus - can be made using a myriad of evidence to make informed decisions and, therefore, these indicators can offer information relevant to assessing strategies. There is also scope for indicators to be developed and analysed further to increase their utility in this respect, for example, by calculating the associated costs. One of the advantages offered by analysing hospital administrative data – that it is linked, in many countries, to funding frameworks – remains largely unexploited. Further, despite countries assessing the value of technology by its cost-effectiveness, such analysis is limited. “There is no full economic evaluation of the burden related to the implementation of various methods... Obtaining such information is usually a secondary objective of the published studies, and no details are available on how the calculations were performed” ⁵⁹. The WHO noted that the cost-benefit analysis of patient safety measures “could provide an impetus and help countries make the necessary investments” (p86). ⁵¹⁵

11.6.3.7. Case findings

The most consistently recommended use of administrative data are for case finding, perhaps representing a ‘safe option’ as it does not introduce any significant incentives to game the indicators. In a paper outlining how the AHRQ PSIs could be applied to the NHS, Bottle and Aylin suggested that the indicators would be a useful tool to assist clinical audit towards potential issues. ¹¹² Similarly, Miller and Zhan have highlighted the appropriate use of the PSIs as institutional case-finding tools aimed at internal quality improvement as opposed to use for directly comparing individual institutions, especially in public reports that identify individual hospitals. ²⁴ In a paper setting out the limitations of hospital mortality rates, Lilford and Pronovost cautioned against the use of such indicators as the basis for sanctions or rewards but did concede that there was value in using such indicators to flag where further investigation might be needed. ⁴⁹

11.6.3.8. As part of measurement portfolio

A full picture of quality and safety would not be possible from any one data source. ⁵²⁸ In particular, for many uses, indicators are not detailed enough when used alone ⁷³ and so “caution should be exercised when using administrative data, exclusively, to report quality and safety outcomes”. ¹¹¹ As well as a plurality of indicators, there are advantages in having a range of types of indicators; recall the discussion from earlier that, for instance, process of care measures are more positively influenced by feedback than outcome of care measures but the latter more intrinsically interesting to most stakeholders (with different data sources being more adept to being used for either outcomes or process measures). ⁵²⁹ As such, rather than conclude spuriously on which is the best
application, I instead recommend that any use of such indicators by policy-makers is, where possible, as part of a broader measurement portfolio. This should also involve patient input and experiences, perhaps in the form of patient reported outcome measures.

### 11.7. Conclusion

This project showed that specialty-specific standardised mortality rates can reveal significant differences in performance both between providers and over time. Moreover, I was able to show that routinely-collected hospital administrative data can offer a far broader understanding of quality and safety than solely mortality rates.

While I set out a range of limitations to both this and existing research that has used administrative data to evaluate quality and safety of healthcare, researchers in this area should not be too bashful about limitations in terms of validity. A paper setting out the limitations of HSMR concluded that “quality of care should remain innocent until proved guilty”, although this view has been subsequently criticised.\(^{512}\) Indeed there is no perfect measure of quality and so, if this maxim was held, then indolence would prevail.

This project itself has shown that using indicators based on administrative data, albeit in this speciality-specific context and applied in a way that overcomes key limitations, can focus government, senior doctors and the public alike on areas of likely substandard care.\(^{530}\) To return to the similarly-themed fable and idiom introduced at the start of this thesis, if such research continues and some of the areas for further work are addressed then the blind men will have a better chance of identifying their subject and the issue of patient safety will no longer remain the elephant in the room.
References

59. Michel P. Strengths and weaknesses of available methods for assessing the nature and scale of harm caused by the health system: literature review; 2003.
82. McDonald KM, Romano PS, Geppert JD, et al. Measures of Patient Safety Based on Hospital Administrative Data - The Patient Safety Indicators. (Technical Reviews, No. 5.). Rockville (MD): Agency for Healthcare Research and Quality (USA); 2002.


160. Queensland Health: Clinical Practice Improvement Centre. Acute Stroke Care Indicators 30/06/2007 (Version 1.7); 2007.


Part IV: Discussion, recommendation and appendices


258. Escarce JJ, Jain AK, Rogowski J. Hospital competition, managed care, and mortality after hospitalization for medical conditions: Evidence from three states. Medical Care Research and Review 2006;63:112S-40S.


312. Williams JY. Inventory of studies on the accuracy of Canadian health administrative databases. Toronto, Ontario; 1996.
314. Canadian Institute for Health Information. Data quality of the discharge abstract database following the first-year implementation of ICD-10-CA/CCI. Final report.; 2004 September.


370. Martinez Marco EA, Aranaz Andres J. Is there a relationship between hospital readmission and quality of care?. [Spanish]
383. Carvalho ML, Silver LD. Reliability of the certification of the basic causes of neonatal deaths: implications for the study of preventable mortality. [Portuguese]

407. Reilly AF, Reilly PM. Medical injuries can increase mortality risk and incur additional costs. Evidence-Based Healthcare 2004;8:60-2.
Part IV: Discussion, recommendation and appendices

447. Gutierrez Saucedo ME, Hernandez Herrera RJ, Luna Garcia SA, Flores Santos R, Alcala Galvan LG, Martinez Gaytan V. Perinatal mortality at Hospital de Ginecoobstetricia no. 23 of Monterrey, Nuevo Leon, 2002-2006 period. [Spanish]
451. Maass C, Schleiz W, Weyermann M, Drosler SE. Are hospital administrative data suitable for external quality assurance? Comparison of quality indicators based on separate statutory data collections (BQS) and hospital administrative data. [German]

451. Maass C, Schleiz W, Weyermann M, Drosler SE. Are hospital administrative data suitable for external quality assurance? Comparison of quality indicators based on separate statutory data collections (BQS) and hospital administrative data. [German]

457. Gerbier S, Bouzbid S, Pradat E, et al. Use of the French medico-administrative database (PMSI) to detect nosocomial infections in the University hospital of Lyon. [French]


525. HFMA. NHS Services, Seven Days a Week Forum: Costing seven day services - The financial implications of seven day services for acute emergency and urgent services and supporting diagnostics. 2014.
528. Thomson R. Adverse events reporting in English hospital statistics: Data relevant to patient safety should not be presented alone and out of context. BMJ 2004;329:857-.
Appendices

Appendix A. Outputs from the project

Publications related to the project


Other outputs


Conference abstracts related to the project


## Appendix B. Details of previous literature reviews of quality and safety indicators

<table>
<thead>
<tr>
<th>Review</th>
<th>Pub Date</th>
<th>Description / Aims</th>
<th>Methods</th>
<th>Search terms</th>
<th>Resources</th>
<th>Inclusion / Exclusion criteria</th>
<th>Research identified</th>
<th>Indicator shortlisting</th>
<th>Final indicators</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>International Quality Improvement Programme (IQIP)</strong></td>
<td>Since 1985 in US and 1991 internationally</td>
<td>Publish topic specific indicators that can be used for benchmarking performance</td>
<td>The measures were reviewed and modified following, firstly, feedback from subject-matter experts and then feedback and data from the first 6 to 12 months of use.</td>
<td>Tailored to the specific topic/procedure&lt;sup&gt;8&lt;/sup&gt;</td>
<td>PubMed. National sites (e.g., NPSA, Joint Commission) and Departments and Ministries of Health in the US and other countries as well as any of the accrediting agencies.</td>
<td></td>
<td></td>
<td>Nearly 700 indicators, covering four care settings: acute, psychiatric (behaviour health) care, long term care, and home care.</td>
<td></td>
</tr>
<tr>
<td><strong>HCUP QIs</strong></td>
<td>1994</td>
<td>Meet the short-term needs for information on health care quality using standardised, user-friendly and existing sources of data</td>
<td>Described in Ball et al (1998)&lt;sup&gt;103&lt;/sup&gt;</td>
<td>Not available</td>
<td>Not available</td>
<td>Not available</td>
<td></td>
<td>HCUP I consisted of 33 quality measures</td>
<td></td>
</tr>
<tr>
<td><strong>Joint Commission’s National Hospital Quality Measures</strong></td>
<td>April 2009 (Version 2.6b) (based on work since 1999)</td>
<td>To construct quality measure set for commissioners and providers of hospital care to publicly report their activities.</td>
<td>Not stated (and unavailable on request)</td>
<td>-</td>
<td>The ORYX&lt;sup&gt;®&lt;/sup&gt; initiative, which became operational in March 1999, has over 8,000 disparate measures.</td>
<td></td>
<td></td>
<td>The Centers for Medicare &amp; Medicaid Services (CMS) chose 10 starter set measures and 11 additional measures from a consensus-derived set of 39 measures.</td>
<td></td>
</tr>
</tbody>
</table>

---

<sup>8</sup> IQIP information from KarolW@mhaonline.org
### Part IV: Discussion, recommendation and appendices

<table>
<thead>
<tr>
<th>Review</th>
<th>Pub Date</th>
<th>Description / Aims</th>
<th>Methods</th>
<th>Search terms</th>
<th>Resources</th>
<th>Inclusion / Exclusion criteria</th>
<th>Research identified</th>
<th>Indicator shortlisting</th>
<th>Final indicators</th>
</tr>
</thead>
<tbody>
<tr>
<td>AHRQ QIs (refinement of HCUP QIs)</td>
<td>2001</td>
<td>To identify Patient Safety Indicators that could be used using ICD-9. They are based on the HCUP QIs.</td>
<td>Steps: 1. Defined the concepts and the evaluation framework; 2. Search the Literature; 3. Develop a Candidate List of PSIs; 4. Review PSIs; 5. Evaluate the PSIs</td>
<td>Medical Subject Heading (MeSH) terms “hospital, statistic, and methods” and “quality indicators”</td>
<td>In addition, electronic searches for articles published before February 2002, followed by hand search and Author search for Iezzoni, given her work on CSP</td>
<td>Inclusion criteria: quality indicator; novel indicator; and based on administrative data</td>
<td>PQI: 2,600 articles since 1994 of which 181 articles provided information on potential quality indicators based on administrative data. 27 met full criteria</td>
<td>Semi-structured interviews, evaluation framework (7 aspects: face validity, precision, minimum bias, construct validity, fosters real quality improvement, application)</td>
<td>28 provider level and 4 area level Inpatient Quality Indicators (IQIs); Patient Safety Indicators; Paediatric Quality Indicators; and Prevention Quality Indicators.</td>
</tr>
<tr>
<td>Millar et al 2004, [OECD Health Care Quality Indicator (HCQI) project]</td>
<td>2004</td>
<td>Develop PSIs to cover: hospital-acquired infections; sentinel events; operative and postoperative complications; obstetrics; and other care-related AEs.</td>
<td>Review of existing routine data-based PSIs in OECD countries. PSIs selected according to indicator importance, scientific soundness and feasibility.</td>
<td>n/a</td>
<td>Previous indicator searches</td>
<td></td>
<td>7 indicator sets: AHRQ PSIs; AHRQ/CIHI PSIs; ACSQ; Complication Screening Programme (BIH, Israel), JCAHO Infection Control; JCAHO Sentinel Events.</td>
<td>Expert panel</td>
<td>21 indicators, from total of 59</td>
</tr>
<tr>
<td>SIMPATIE</td>
<td>2006</td>
<td>Described in Kristensen et al (2007)101</td>
<td>The development of indicators was based on an adjusted method described in Mainz (2003)102</td>
<td>An extensive literature search was initiated using the search terms: “Patient safety”, “Indicator”, “Risk” and “Harm”.</td>
<td>PubMed and <a href="http://scholar.google.dk/">http://scholar.google.dk/</a> were searched. 9</td>
<td>Search terms: “Test**”, “Usage”/”Use”, “Apply” and “Valid**” were used to assess if indicators previously evaluated in clinical setting.</td>
<td>Indicators or indicator programmes of the organisations mentioned in ‘resources’ were reviewed to determine whether they were suitable for characterisation.</td>
<td>3 aspects: relevance; validity and reliability; feasibility</td>
<td>42 indicators with 24 recommended in all or part of Europe</td>
</tr>
</tbody>
</table>

9 Indicator programmes were reviewed Australian Council for Safety and Quality in Health Care (ACSQHC); Agency for Healthcare Research and Quality (AHRQ); The Good Medical Department, Denmark (DGMA); International Compendium of Health Indicators (WHO, OECD, Eurostat and ECHIM; Institute of Healthcare Improvements (IHI); Joint Commission on accreditation in Health Care (JCAHO); Nordic Indicators (NI); Performance Assessment Tool for Quality improvement in Hospitals (PATH); The Danish National Indicator Project (NIP); Performance Indicators on Patient Safety and effectiveness for Dutch Hospitals.
<table>
<thead>
<tr>
<th>Review</th>
<th>Pub Date</th>
<th>Description / Aims</th>
<th>Methods</th>
<th>Search terms</th>
<th>Resources</th>
<th>Inclusion / Exclusion criteria</th>
<th>Research identified</th>
<th>Indicator shortlisting</th>
<th>Final indicators</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>NIP (Denmark)</strong></td>
<td></td>
<td>Perform focused literature searches concerning the existence and experience with measurable parameters. Simultaneously search relevant international indicator libraries</td>
<td></td>
<td>e.g. the AHRQ, National Quality Measures Clearinghouse, indicators from professional society websites (mainly US and UK) and indicators from international agencies, such as OECD, Joint Commission etc.)</td>
<td></td>
<td></td>
<td>The indicators were developed, tailored to the chosen disease area, based on a search of scientific evidence or, in its absence, the decided by expert consensus.</td>
<td>For each disease about 6-10 indicators were determined.</td>
<td></td>
</tr>
<tr>
<td><strong>National Indicators of Safety and Quality Project (Australia)</strong></td>
<td></td>
<td>The Australian Council for Safety and Quality in Health Care was established in January 2000 to improve the safety and quality of health care provision in Australia. Report on: suitability of OECD indicators for Australian data; measuring and reporting mortality; possible primary care indicators recommending indicators to support monitoring and improvement in safety and quality of health care.</td>
<td>General search of patient safety literature</td>
<td></td>
<td></td>
<td>230 papers from both health (152 papers) and non-health literature (78 papers) that are case studies, literature reviews, research studies or guideline documents containing data.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Patient safety indicators: A systematic review of the literature Carmen Tsang, Paul Aylin, William Palmer</strong></td>
<td>Oct 2008 (11/07 &amp; 03/08)</td>
<td>review the literature on patient safety indicators developed from routinely collected hospital data</td>
<td>Systematic Literature review</td>
<td>patient safety OR hospital safety OR patient safety indicator$ OR safety improvement$ AND adverse event$ OR medical error$ OR sentinel event$ OR healthcare associated injury OR healthcare associated injuries OR iatrogenic disease$ OR preventable complication$</td>
<td>Medline (1950 to present) and Embase (1980 to present); Inc. grey literature &amp; Government websites (Aus, Can, Den, UK, US)</td>
<td>Exclusion: - Studies outside of secondary care; and - Studies primarily on specific diseases, diagnoses or treatments; or - Studies evaluating teaching or research tools; and - Articles not in English.</td>
<td>Total 1517 citations (1295 databases, 158 from reference lists, 43 from websites) 726 non-relevant, 244 duplications, 75 non-English, 2 not available, 2 non-human Of 468 fully abstracted, 277 either Strategies or Background and 91 articles on PSIs</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Notes:** ^ The search terms for health related literature were based on the standard medical subject headings (MeSH) and include any or all of the following terms:
- Manpower, doctors, medical staff, nursing staff, health personnel, staffing, personnel staffing, scheduling, personnel administration, workload
- Nursing staff, hospital/supply, distribution
- Quality assurance
- Hospital safety management
- Patient safety, patient incidents, patient adverse health effects, clinical risk, quality management, quality care, risk management, quality of healthcare, quality indicators, outcome assessment/methods, hospital/standards, accident prevention, safety management, adverse events, adverse drug reactions
- Staff* rostering, staff fatigue, staff supervision, work place harassment, staff dissatisfaction, clinical governance, staff competency. # The term 'out-of-hours' was not searchable in the majority of databases used for this review. As a general rule, the term 'of' is not searchable in a database as
Part IV: Discussion, recommendation and appendices

It is considered a ‘stop’ word that occurs too frequently to make search results relevant. ‘Out-of-hours’ is found in the searching related to workload, scheduling and working hours. * The term “staff” will include doctors, nurses, midwives, managers, surgeons, anaesthetists, physicians, allied health professionals.
Appendix C: Additional websites searched

The list below was adapted from a list used by Tsang and colleagues in their review of general patient safety indicators.66

<table>
<thead>
<tr>
<th>Organisation</th>
<th>URL</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>International</strong></td>
<td></td>
</tr>
<tr>
<td>European Commission Health Indicators Monitoring</td>
<td><a href="http://www.echim.org/index.html">http://www.echim.org/index.html</a></td>
</tr>
<tr>
<td>European Observatory on Health Systems and Policies</td>
<td><a href="http://www.euro.who.int/observatory">http://www.euro.who.int/observatory</a></td>
</tr>
<tr>
<td>Organization for Economic Co-operation and Development</td>
<td><a href="http://www.oecd.org/home/0,2987,en_2649_201185_1_1_1_1_1_00.html">http://www.oecd.org/home/0,2987,en_2649_201185_1_1_1_1_1_00.html</a></td>
</tr>
<tr>
<td>Safety Improvement for Patients In Europe Project</td>
<td><a href="http://www.simpatie.org/Main">http://www.simpatie.org/Main</a></td>
</tr>
<tr>
<td>The Commonwealth Fund</td>
<td><a href="http://www.commonwealthfund.org/">www.commonwealthfund.org/</a></td>
</tr>
<tr>
<td>The Joint Commission International Center for Patient Safety</td>
<td><a href="http://www.jointcommissioninternational.org/">http://www.jointcommissioninternational.org/</a></td>
</tr>
<tr>
<td>WHO Collaborating Centre for Patient Safety Solutions</td>
<td><a href="http://ccforpatientsafety.org/">http://ccforpatientsafety.org/</a></td>
</tr>
<tr>
<td>World Health Organization Regional Office for Europe</td>
<td><a href="http://www.euro.who.int/">http://www.euro.who.int/</a></td>
</tr>
<tr>
<td><strong>Australia</strong></td>
<td></td>
</tr>
<tr>
<td>Australian Patient Safety Foundation</td>
<td><a href="http://www.apsf.net.au/">http://www.apsf.net.au/</a></td>
</tr>
<tr>
<td>Royal Australian College of General Practice</td>
<td><a href="http://www.racgp.org.au/">http://www.racgp.org.au/</a></td>
</tr>
<tr>
<td>Royal Australian College of Physicians</td>
<td><a href="http://www.racp.edu.au/">http://www.racp.edu.au/</a></td>
</tr>
<tr>
<td>Royal College of Nursing</td>
<td><a href="http://www.rcna.org.au/">http://www.rcna.org.au/</a></td>
</tr>
<tr>
<td><strong>Canada</strong></td>
<td></td>
</tr>
<tr>
<td>Accreditation Canada</td>
<td><a href="http://www.accreditationcanada.ca/default.aspx">http://www.accreditationcanada.ca/default.aspx</a></td>
</tr>
<tr>
<td>Canadian Institute for Health Information</td>
<td><a href="http://secure.cii.ca/cihiweb/dispPage.jsp?cw_page=home_e">http://secure.cii.ca/cihiweb/dispPage.jsp?cw_page=home_e</a></td>
</tr>
<tr>
<td>Canadian Patient Safety Institute</td>
<td><a href="http://www.patientsafetyinstitute.ca/index.html">http://www.patientsafetyinstitute.ca/index.html</a></td>
</tr>
<tr>
<td><strong>Denmark</strong></td>
<td></td>
</tr>
<tr>
<td>The Danish Society for Patient Safety</td>
<td><a href="http://www.patientsikkerhed.dk/en/about_the_danish_society_for_patient_safety/about_the_society/">http://www.patientsikkerhed.dk/en/about_the_danish_society_for_patient_safety/about_the_society/</a></td>
</tr>
<tr>
<td><strong>United Kingdom</strong></td>
<td></td>
</tr>
<tr>
<td>Centre for Reviews and Dissemination, University of York</td>
<td><a href="http://www.crd.york.ac.uk/">http://www.crd.york.ac.uk/</a></td>
</tr>
<tr>
<td>Clinical and Health Outcomes Knowledge Base, National Centre for Health Outcomes Development</td>
<td><a href="http://www.nchod.nhs.uk/">http://www.nchod.nhs.uk/</a></td>
</tr>
<tr>
<td>National Audit Office</td>
<td><a href="http://www.nao.org.uk/">http://www.nao.org.uk/</a></td>
</tr>
<tr>
<td>National Patient Safety Agency</td>
<td><a href="http://www.npsa.nhs.uk/">http://www.npsa.nhs.uk/</a></td>
</tr>
<tr>
<td>Patient Safety Research Portfolio, University of Birmingham</td>
<td><a href="http://www.pcpoh.bham.ac.uk/publichealth/psrp/index.shtml">http://www.pcpoh.bham.ac.uk/publichealth/psrp/index.shtml</a></td>
</tr>
<tr>
<td>Royal College of Obstetricians and Gynaecologists</td>
<td><a href="http://www.rcog.org.uk">http://www.rcog.org.uk</a></td>
</tr>
<tr>
<td>Royal College of Midwives</td>
<td><a href="http://www.rcm.org.uk">http://www.rcm.org.uk</a></td>
</tr>
<tr>
<td><strong>United States of America</strong></td>
<td></td>
</tr>
<tr>
<td>Agency for Healthcare Research and Quality</td>
<td><a href="http://www.ahrq.gov/">http://www.ahrq.gov/</a></td>
</tr>
<tr>
<td>Center on Health Policy/Center on Primary Care and Outcomes Research, Stanford University</td>
<td><a href="http://healthpolicy.stanford.edu/">http://healthpolicy.stanford.edu/</a></td>
</tr>
<tr>
<td>Centers for Disease Control and Prevention</td>
<td><a href="http://www.cdc.gov/ncidod/hip/default.htm">http://www.cdc.gov/ncidod/hip/default.htm</a></td>
</tr>
<tr>
<td>Healthcare Cost and Utilization Project</td>
<td><a href="http://hcupnet.ahrq.gov/">http://hcupnet.ahrq.gov/</a></td>
</tr>
<tr>
<td>Institute for Healthcare Improvement</td>
<td><a href="http://www.ihi.org/">http://www.ihi.org/</a></td>
</tr>
<tr>
<td>Institute of Medicine</td>
<td><a href="http://www.iom.edu/">http://www.iom.edu/</a></td>
</tr>
<tr>
<td>National Coalition on Health Care</td>
<td><a href="http://www.nchc.org/">http://www.nchc.org/</a></td>
</tr>
<tr>
<td>National Coordinating Council for Medication Error Reporting and Prevention</td>
<td><a href="http://www.nccmerp.org/">http://www.nccmerp.org/</a></td>
</tr>
<tr>
<td>National Institutes of Health</td>
<td><a href="http://www.nih.gov/">http://www.nih.gov/</a></td>
</tr>
<tr>
<td>National Committee for Quality Assurance</td>
<td><a href="http://www.ncqa.org/">http://www.ncqa.org/</a></td>
</tr>
<tr>
<td>National Patient Safety Foundation</td>
<td><a href="http://www.npsf.org/">http://www.npsf.org/</a></td>
</tr>
<tr>
<td>----------------------------------------</td>
<td>----------------------</td>
</tr>
<tr>
<td>Quality Interagency Coordinating Task Force</td>
<td><a href="http://www.quic.gov/">http://www.quic.gov/</a></td>
</tr>
<tr>
<td>The Joint Commission</td>
<td><a href="http://www.jointcommission.org/">http://www.jointcommission.org/</a></td>
</tr>
<tr>
<td>National Center for Patient Safety, Department of Veterans Affairs</td>
<td><a href="http://www.patientsafety.gov/">http://www.patientsafety.gov/</a></td>
</tr>
</tbody>
</table>
## Appendix D: Stroke indicators

<table>
<thead>
<tr>
<th>Indicator</th>
<th>Definition (example)</th>
<th>Reference (example)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Urgent brain scan for stroke patients</strong></td>
<td>Proportion of stroke patients who have a brain scan (CT or MRI) performed on the day of (or by following day after) admission</td>
<td>Queensland Government&lt;sup&gt;532&lt;/sup&gt;</td>
</tr>
<tr>
<td><strong>Thrombolysis</strong></td>
<td>Proportion of patients receiving thrombolysis treatment</td>
<td>American Academy of Neurology (2010)&lt;sup&gt;200&lt;/sup&gt;</td>
</tr>
<tr>
<td><strong>Emergency hospital transfers</strong></td>
<td>All patients with suspected acute stroke are immediately transferred by ambulance to a receiving hospital providing hyper-acute stroke services</td>
<td>Kind (2007)&lt;sup&gt;272&lt;/sup&gt;</td>
</tr>
<tr>
<td><strong>Urinary Tract Infections (UTIs)</strong></td>
<td>Proportion of stroke admissions with secondary diagnoses of urinary tract infection</td>
<td>Laudicella (2008)&lt;sup&gt;221&lt;/sup&gt;</td>
</tr>
<tr>
<td><strong>Pneumonia due to swallowing problems</strong></td>
<td>Proportion of stroke admissions with pneumonitis due to food or vomiting (a consequence of swallowing problems)</td>
<td>Laudicella (2008)&lt;sup&gt;221&lt;/sup&gt;</td>
</tr>
<tr>
<td><strong>Lower respiratory tract infection (LRTI)</strong></td>
<td>Proportion of stroke admissions with unspecified acute lower respiratory tract infection</td>
<td>Laudicella (2008)&lt;sup&gt;221&lt;/sup&gt;</td>
</tr>
<tr>
<td><strong>Deep Vein Thrombosis</strong></td>
<td>Proportion of stroke admissions with secondary diagnoses of deep vein thrombosis</td>
<td>Cnesullo (2008)&lt;sup&gt;243&lt;/sup&gt;</td>
</tr>
<tr>
<td><strong>Pulmonary Embolism</strong></td>
<td>Proportion of stroke patients with documented pulmonary embolism occurring during admission</td>
<td>Field (2004)&lt;sup&gt;219&lt;/sup&gt;</td>
</tr>
<tr>
<td><strong>Clostridium difficile</strong></td>
<td>Proportion of stroke admissions with Enterocolitis due to Clostridium difficile</td>
<td>Laudicella (2008)&lt;sup&gt;221&lt;/sup&gt;</td>
</tr>
<tr>
<td><strong>Depression and other mental health problems following stroke</strong></td>
<td>Diagnosis of depression and other mental health diagnosis in the first 3 years post-stroke</td>
<td>Williams (2004)&lt;sup&gt;228&lt;/sup&gt;</td>
</tr>
<tr>
<td>Measure</td>
<td>Description</td>
<td>Source</td>
</tr>
<tr>
<td>----------------------------------------------</td>
<td>-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------</td>
<td>--------------</td>
</tr>
<tr>
<td><strong>Inpatient rehabilitation services</strong></td>
<td>Proportion of stroke admissions with recorded use of rehabilitation services, defined as reimbursement claims for physical therapy, occupational therapy, speech therapy, or a combination of these therapies</td>
<td>Kapral (2002)</td>
</tr>
<tr>
<td><strong>Warfarin use</strong></td>
<td>Warfarin use as proxied by outpatient claims for prothrombin time tests</td>
<td>Smith (2005)</td>
</tr>
<tr>
<td><strong>Length of stay</strong></td>
<td>Length of hospitalisation</td>
<td>Zhu (2009)</td>
</tr>
<tr>
<td><strong>In-hospital mortality within 30 days of admission</strong></td>
<td>The standardised proportion of stroke patients who die in-hospital within 30 days of admission</td>
<td>Queensland Government</td>
</tr>
<tr>
<td><strong>All-location mortality within 28 days of admission</strong></td>
<td>Proportion of stroke patients who die within 28 days of admission</td>
<td>Pajunen (2005)</td>
</tr>
<tr>
<td><strong>Craniotomy: mortality rate</strong></td>
<td>Number of deaths per 100 discharges with a code for craniotomy, with and without comorbidities and complications.</td>
<td>AHRQ (2013)</td>
</tr>
<tr>
<td><strong>Return to usual place of residence</strong></td>
<td>Proportion of stroke patients returning to their usual place of residence following hospital treatment within 56 days</td>
<td>Qureshi (2005)</td>
</tr>
<tr>
<td><strong>Emergency readmission (all cause) to hospital</strong></td>
<td>Percentage of patients of all ages with emergency readmission to any hospital within 27 days (inclusive) of the last, previous discharge from hospital after admission with a stroke</td>
<td>Rosato (2006)</td>
</tr>
<tr>
<td><strong>Recurrent stroke within 28 days</strong></td>
<td>Percentage of patients of all ages with emergency readmission for stroke to any hospital within 27 days (inclusive) of the last, previous discharge from hospital after admission with a stroke</td>
<td>Johansen (2006)</td>
</tr>
<tr>
<td><strong>Readmission for incontinence</strong></td>
<td>Percentage of patients readmitted to any hospital within 27 days with diagnosis for incontinence</td>
<td>Queensland Government</td>
</tr>
</tbody>
</table>
### Appendix E: Obstetrics indicators

<table>
<thead>
<tr>
<th>Indicator</th>
<th>Indicator type</th>
<th>Detail</th>
<th>Inclusions or exclusions</th>
<th>Alternatives</th>
<th>Indicator sets, primary (other)</th>
<th>Key articles</th>
<th>Validation</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>For mother</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>M1. Non-elective readmissions</td>
<td>Outcome</td>
<td>Within 42 days of the start of a delivery episode</td>
<td>Excludes readmissions of less than a day</td>
<td>1. Different time periods: any postnatal admission of mother (NCHOD); 6 months (Jimenez-Puente 2002); 90 day [400]; 30 day [423] (Tseng 2010); 15 days (Tseng 2010); 7 or 14 days (Kotagal 1999) 2. Readmissions for PPH [440] 3. Readmissions within 28 days [535]</td>
<td>CQC, NPIC</td>
<td>390,398, 400,423,536</td>
<td></td>
</tr>
<tr>
<td>M2. Puerperal sepsis</td>
<td>Outcome</td>
<td>ICD-10 diagnosis code of O85 at any point during delivery spell within 42 days of the start of a delivery spell</td>
<td>Includes readmission within 42 days of the start of the delivery spell.</td>
<td>1. Puerperal sepsis and other puerperal infection (ICD-10 O85 and O86) or specified puerperal infection (excludes O86.4) 2. Chorioamnionitis (JCAHO) 3. Uterine infection (Korst 2005) 4. Wound complications (AHRQ)</td>
<td>CQC</td>
<td>427, 392, 415,428,435,444,452</td>
<td></td>
</tr>
<tr>
<td>M3. Obstetric Trauma</td>
<td>Outcome /Proxy outcome (procedures)</td>
<td>Stratified by delivery type (VB with instrument; VB without instrument; CS) Third or fourth degree tear in dx or procedure codes</td>
<td>Includes uterine rupture, fracture of pelvis, including coccyx, laceration or haematoma of cervix, vagina, vulva, perineum and anus.</td>
<td>1. Trauma to perineal (PERISTAT) 2. Cases of obstetric trauma (third or fourth degree lacerations) (JCAHO) 3. Selected primipara with an intact perineum or unsutured perineal tear 4. Surgical repair of perineum (ACHS) 5. Aggregation of w/ and w/o</td>
<td>AHRQ, NPCI, JCAHO, NOGA, RANZCAG/ACHS, SimPatIE, DNBH, RCOG</td>
<td>83,375,391,393,401,404,407, 409,416,427,435,437,441,444,451,45 2,435,456,461</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Indicator</td>
<td>Indicator type</td>
<td>Detail</td>
<td>Inclusions or exclusions</td>
<td>Alternatives</td>
<td>Indicator sets, primary (other)</td>
<td>Key articles</td>
<td>Validation</td>
</tr>
<tr>
<td>---------------------------------</td>
<td>----------------</td>
<td>------------------------------------------------------------------------</td>
<td>--------------------------------------------------------------------------------------------</td>
<td>------------------------------------------------------------------------------</td>
<td>--------------------------------</td>
<td>--------------</td>
<td>------------</td>
</tr>
<tr>
<td>M4. Episiotomy</td>
<td>Proxy outcome</td>
<td>Perineal trauma and episiotomy rates (NCHOD)</td>
<td>Instrument (OECD)</td>
<td>1. Perineal trauma and episiotomy rates (NCHOD)</td>
<td>JCAHO (future measures), NOGA</td>
<td>Other 392</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Separated into selected primipara undergoing episiotomy AND sustaining a perineal tear during VB or with NO tear</td>
<td></td>
<td>2. Separated into selected primipara undergoing episiotomy AND sustaining a perineal tear during VB or with NO tear</td>
<td></td>
<td></td>
<td>418</td>
</tr>
<tr>
<td>M5. Uterine rupture</td>
<td>Outcome</td>
<td>Post-natal urinary or faecal incontinence</td>
<td>Discharges with disposition of “deceased”</td>
<td>1. Faecal incontinence (PERISTAT)</td>
<td>NPIC</td>
<td></td>
<td></td>
</tr>
<tr>
<td>M6. Incontinence</td>
<td>Outcome</td>
<td>Post-natal urinary or faecal incontinence</td>
<td>Discharges with disposition of “deceased”</td>
<td>1. Faecal incontinence (PERISTAT)</td>
<td>NCHOD</td>
<td></td>
<td></td>
</tr>
<tr>
<td>M7. Mortality</td>
<td>Outcome</td>
<td>Limited to only delivery admissions</td>
<td>Discharges with disposition of “deceased”</td>
<td>1. Average LOS (OECD)</td>
<td>AHRQ (death in low-mortality DRGs), OBCQID, NCHOD, OECD, PERISTAT</td>
<td>391,442,446,454</td>
<td>Core measure in PERISTAT</td>
</tr>
<tr>
<td>M8. Long Length of stay</td>
<td>Outcome</td>
<td>Discharges with dx code for anaesthesia complications in any dx field</td>
<td>Discharges with dx code for anaesthesia complications in any dx field</td>
<td>1. Average LOS (OECD)</td>
<td>NPIC</td>
<td></td>
<td>389,390,398,461</td>
</tr>
<tr>
<td>M9. Complications of anaesthesia</td>
<td>Outcome</td>
<td>Discharges with dx code for anaesthesia complications in any dx field</td>
<td>Discharges with dx code for anaesthesia complications in any dx field</td>
<td>1. Average LOS (OECD)</td>
<td>AHRQ</td>
<td></td>
<td>391,395,433,444</td>
</tr>
<tr>
<td>M10. Foreign body left during procedure</td>
<td>Outcome</td>
<td>Discharges with dx code for foreign body left in during procedure in any dx field</td>
<td>Discharges with dx code for foreign body left in during procedure in any dx field</td>
<td>1. Average LOS (OECD)</td>
<td>AHRQ</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Indicator</td>
<td>Indicator type</td>
<td>Detail</td>
<td>Inclusions or exclusions</td>
<td>Alternatives</td>
<td>Indicator sets, primary (other)</td>
<td>Key articles</td>
<td>Validation</td>
</tr>
<tr>
<td>-----------</td>
<td>----------------</td>
<td>--------</td>
<td>--------------------------</td>
<td>--------------</td>
<td>---------------------------------</td>
<td>--------------</td>
<td>------------</td>
</tr>
<tr>
<td>M11. Post-partum haemorrhage (PPH)</td>
<td>Outcome</td>
<td>Diagnosis for postoperative haemorrhage/haematoma or related procedures</td>
<td>Procedure codes for control of haemorrhage must occur within one day after birth</td>
<td>1. Obstetric haemorrhage (ACOG) 2. PPH or transfusion (RANZCOG/ACHS) 3. Wider measure of PPH morbidity</td>
<td>AHRQ (postoperative haemorrhage or hematoma), NCHOD</td>
<td>391,392,405,427,435,440</td>
<td>Korst et al (2005) recommended</td>
</tr>
<tr>
<td>M12. Transfusion</td>
<td>Proxy outcome</td>
<td>Women who give birth vaginally who receive a blood transfusion during the same admission.</td>
<td></td>
<td>1. Transfusions (Korst et al 2005) PPH or transfusion (RANZCOG/ACHS) 2. Transfusions for CS (ACHS)</td>
<td>ACHS</td>
<td>389,405,427,444</td>
<td></td>
</tr>
<tr>
<td>M13. Transfusion reaction</td>
<td>Outcome</td>
<td></td>
<td></td>
<td></td>
<td>AHRQ-derived</td>
<td></td>
<td></td>
</tr>
<tr>
<td>M14. Episiotomy</td>
<td>Proxy outcome</td>
<td></td>
<td>1. Perineal trauma and episiotomy rates (NCHOD) 2. Separated into selected primipara undergoing episiotomy AND sustaining a perineal tear during VB/NO tear</td>
<td>JCAHO (future measures), NOGA</td>
<td>418,436,452,456</td>
<td></td>
<td></td>
</tr>
<tr>
<td>M15. Maternal transfer to perinatal centre</td>
<td>Outcome</td>
<td></td>
<td>1. Transfer to ICU (NCHOD)</td>
<td>JCAHO (future measures)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>M16. Eclampsia</td>
<td>Process</td>
<td>ICD-10 O15</td>
<td></td>
<td>OBCQID, NCHOD, ACOG</td>
<td>391</td>
<td></td>
<td></td>
</tr>
<tr>
<td>M17. Severe maternal morbidity</td>
<td>Composite</td>
<td></td>
<td>1. Adverse Outcome Index (Walker 2010) 2. Maternal composite (Srinivas 2010) 3. Weighted adverse outcome score</td>
<td>PERISTAT</td>
<td>393,410,424,429,430,442,443,446,449,453,454,458,459</td>
<td>Recommended for further development (PERISTAT) Requires extensive data collected and</td>
<td></td>
</tr>
<tr>
<td>Indicator</td>
<td>Indicator type</td>
<td>Detail</td>
<td>Inclusions or exclusions</td>
<td>Alternatives</td>
<td>Indicator sets, primary (other)</td>
<td>Key articles</td>
<td>Validation</td>
</tr>
<tr>
<td>----------</td>
<td>----------------</td>
<td>--------</td>
<td>--------------------------</td>
<td>--------------</td>
<td>--------------------------------</td>
<td>--------------</td>
<td>------------</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>M19. Cardiac arrest</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>M20. General anaesthesia for CS</td>
<td>Process</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>M21. Pharmacological thromboprophylaxis &amp; CS</td>
<td>Process</td>
<td>Total number of high risk women undergoing caesarean section who receive appropriate pharmacological thromboprophylaxis</td>
<td></td>
<td>1. Thrombosis (Srinivas 2009)</td>
<td>RANZCOG/ACHS</td>
<td>428</td>
<td></td>
</tr>
<tr>
<td>M22. Timely antenatal assessment</td>
<td>Process</td>
<td>Within 10-14 weeks or by 12 weeks</td>
<td></td>
<td></td>
<td>NHS Information Centre</td>
<td>537</td>
<td></td>
</tr>
</tbody>
</table>

**For delivery method**

<table>
<thead>
<tr>
<th>D1. Primary caesarean birth rate</th>
<th>Proxy outcome</th>
<th>Selected primipara only</th>
<th>1. Caesarean section rates (NPIC, NOGA, OECD, CIHI).</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>D2. VBAC</td>
<td>Outcome</td>
<td>vaginal birth after</td>
<td>Any dx of abnormal</td>
<td></td>
<td>AHRQ, JCAHO,</td>
<td>402</td>
<td>Korst et al (2005)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Validation**

risk adjustment (Janakiraman 2010)

83, 391, 392, 461

Goff 2000

RANZCOG/ACHS

428

NHS Information Centre

389, 393, 402, 403, 41
3, 414, 417, 421, 423, 425, 436, 445, 460
Korst et al (2005) recommended

402
Korst et al (2005)
### Part IV: Discussion, recommendation and appendices

<table>
<thead>
<tr>
<th>Indicator</th>
<th>Indicator type</th>
<th>Detail</th>
<th>Inclusions or exclusions</th>
<th>Alternatives</th>
<th>Indicator sets, primary (other)</th>
<th>Key articles</th>
<th>Validation</th>
</tr>
</thead>
<tbody>
<tr>
<td>D3.</td>
<td>Unsuccessful VBAC</td>
<td>Outcome</td>
<td>Cesarean (VBAC) delivery rate, uncomplicated presentation, preterm, foetal death, multiple gestation diagnosis code. Breech procedures codes.</td>
<td></td>
<td>NPIC, RANZCOG/ACHS</td>
<td>JCAHO (future measures), ACOG recommended</td>
<td></td>
</tr>
<tr>
<td>D4.</td>
<td>Failed vacuum or forcep delivery</td>
<td>Outcome</td>
<td></td>
<td>1. CS after failed instrumental delivery (RCOG)</td>
<td></td>
<td>Roberts 2009 RCOG 375,456</td>
<td></td>
</tr>
<tr>
<td>D5.</td>
<td>Emergency caesarean section</td>
<td>Outcome</td>
<td></td>
<td></td>
<td></td>
<td>RCOG 388,397,420,456</td>
<td></td>
</tr>
<tr>
<td>D6.</td>
<td>Late elective VB or CS</td>
<td>Process</td>
<td>At greater or equal to 37 and less than 39 weeks of gestation completed</td>
<td>1. Late CS (Gurol-Urganci 2011) (Benedetti 2009)</td>
<td></td>
<td>JCAHO 411,434</td>
<td></td>
</tr>
<tr>
<td>D7.</td>
<td>Normal births</td>
<td>Composite</td>
<td>A normal delivery is one without induction, without the use of instruments, not by caesarean section and without general, spinal or epidural anaesthetic before or during delivery. Excludes: induction of labour; epidural or spinal; general anaesthetic; forceps or ventouse; CS; episiotomy</td>
<td>1. Operative VB rate (NPIC) 2. Instrumental delivery (OBCQID, Dr Foster) 3. Forceps or ventouse deliveries (NOGA) 4. Births without medical intervention (PERISTAT) 5. Outcome of selected primapara (RANZCOG/ACHS) 6. Indications and/or rate of elective labour (JCAHO future measures) 7. Spontaneous vs instrument VB</td>
<td>BirthChoiceUK, PERISTAT, RCOG 375,389,394,436</td>
<td>Further development recommended (PERISTAT) Instrumental VB recommended by Sibanda 2013</td>
<td></td>
</tr>
<tr>
<td>Indicator</td>
<td>Indicator type</td>
<td>Detail</td>
<td>Inclusions or exclusions</td>
<td>Alternatives</td>
<td>Indicator sets, primary (other)</td>
<td>Key articles</td>
<td>Validation</td>
</tr>
<tr>
<td>-----------</td>
<td>----------------</td>
<td>--------</td>
<td>--------------------------</td>
<td>-------------</td>
<td>-------------------------------</td>
<td>--------------</td>
<td>------------</td>
</tr>
<tr>
<td></td>
<td>Indicator</td>
<td>type</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D8.</td>
<td>Induction rates</td>
<td>Proxy</td>
<td>Deliveries with surgical and/or medical induction of labour</td>
<td>Selected primipar only</td>
<td>ACHS</td>
<td></td>
<td></td>
</tr>
<tr>
<td>D9.</td>
<td>Induction of labour resulting in emergency CS</td>
<td>Outcome</td>
<td></td>
<td></td>
<td>RCOG</td>
<td></td>
<td>375</td>
</tr>
<tr>
<td>For neonate</td>
<td>Non-elective readmissions</td>
<td>Outcome</td>
<td>Within 28 days of birth</td>
<td>Excludes readmissions of less than a day</td>
<td>CQC</td>
<td>412</td>
<td></td>
</tr>
<tr>
<td>N1.</td>
<td>Infections</td>
<td>Outcome</td>
<td>Live babies of greater than or equal to 37 weeks gestational age (GA) born at the reporting hospital who develop a blood and/or cerebrospinal fluid (CSF) infection within 48 hours</td>
<td>1. Babies admitted to the neonatal intensive care unit (NICU) who have a significant blood infection occurring more than 48 hours after birth at any time during their whole admission - disaggregated for those above or below 1000 grams birth weight. (ACHS also) 2. Chorioamnionitis (JCAHO) 3. Neonatal sepsis (VON) 4. Part of wider category of system problems (Gould 2004)</td>
<td>AHRQ (ACHS)</td>
<td>412,415,432</td>
<td>Korst et al (2005) recommended (chorioamnionitis and morbidities) Detailed evaluation of validity and feasibility 415</td>
</tr>
<tr>
<td>N2.</td>
<td>Infections</td>
<td>Outcome</td>
<td>Discharges with code for preterms with</td>
<td>1. Brachial plexus injury (Ford 2007)</td>
<td>AHRQ, OECD, 81,393,401,407,</td>
<td>Janakiraman et al</td>
<td></td>
</tr>
</tbody>
</table>
## Part IV: Discussion, recommendation and appendices

<table>
<thead>
<tr>
<th>Indicator</th>
<th>Indicator type</th>
<th>Detail</th>
<th>Inclusions or exclusions</th>
<th>Alternatives</th>
<th>Indicator sets, primary (other)</th>
<th>Key articles</th>
<th>Validation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Injury to Neonate</td>
<td></td>
<td>birth trauma in any diagnosis field per 1000 live births</td>
<td>subdural or cerebral haemorrhage; infant dx of injury to skeleton or osteogenesis imperfecta</td>
<td>2. Wider definition of trauma (Moczygemba 2010) 3. Brachial plexus injury excluded (SimPatIE)</td>
<td>CIHI, SimPatIE, DNBH</td>
<td>(2010) suggest risk adjustment not well established. Generally performed well in AHRQ validation.</td>
<td></td>
</tr>
<tr>
<td>N5. Iatrogenic pneumothorax</td>
<td>Outcome</td>
<td>dx code for iatrogenic pneumothorax in principle field; any dx code of chest trauma or pleural effusion or selected procedures</td>
<td>Less than 28 days and Birth-weight &lt;500g;</td>
<td></td>
<td></td>
<td>Gould 2004</td>
<td></td>
</tr>
<tr>
<td>N6. Term NICU admissions</td>
<td>Proxy outcome</td>
<td>Percentage of inborn term babies transferred/admitted to a neonatal intensive care nursery or special care nursery facility or reasons of congenital abnormality</td>
<td>1. Neonatal admission to intensive or special care (NCHOD) 2. Neonatal transfer to perinatal centre (JCAHO future measures)</td>
<td>ACHS (JCAHO)</td>
<td>Korst et al (2005) recommended Sibanda (2013) recommended</td>
<td></td>
<td></td>
</tr>
<tr>
<td>N7. Long length of stay</td>
<td>Proxy outcome</td>
<td>Infant stay over 3 days for VB and 5 days for CS</td>
<td></td>
<td>Gould et al 2004</td>
<td>432</td>
<td></td>
<td></td>
</tr>
<tr>
<td>N9. Dehydration</td>
<td>Outcome</td>
<td></td>
<td></td>
<td></td>
<td>Meara 2004</td>
<td>412</td>
<td></td>
</tr>
<tr>
<td>N10. Adverse Outcome</td>
<td>Composite</td>
<td>Selected dx and procedure codes</td>
<td></td>
<td>1. Medical errors (Kanter 2004)</td>
<td>Lain et al (2011)</td>
<td>393,426,430,438</td>
<td></td>
</tr>
<tr>
<td>Indicator</td>
<td>Indicator type</td>
<td>Detail</td>
<td>Inclusions or exclusions</td>
<td>Alternatives</td>
<td>Indicator sets, primary (other)</td>
<td>Key articles</td>
<td>Validation</td>
</tr>
<tr>
<td>----------------------------------</td>
<td>----------------</td>
<td>------------------------------------</td>
<td>--------------------------</td>
<td>-------------------------------------------------------------------------------</td>
<td>--------------------------------</td>
<td>--------------------------------------------------------------------------------</td>
<td>------------</td>
</tr>
<tr>
<td>Indicator</td>
<td></td>
<td>indicating adverse outcome</td>
<td></td>
<td>2. Neonatal composite (Srinivas 2010)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Mother and neonate</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Proportion of births with no</td>
<td>Composite</td>
<td>A “combination” of the above</td>
<td></td>
<td>Include/exclude the individual measures (e.g. mortality) to make a different</td>
<td></td>
<td>1. Ideal delivery (Gregory 2009)</td>
<td>448</td>
</tr>
<tr>
<td>trauma, infection or readmissions</td>
<td></td>
<td></td>
<td></td>
<td>composite measure</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Notes:**


*Selected primipara is defined as: A woman who is 20-34 years of age at the time of giving birth; Giving birth for the first time at greater than 20 weeks of gestation; Singleton pregnancy; Cephalic presentation; At 37 weeks to 41 weeks gestation