Statistical implications of centralised care for estimating neonatal unit mortality rates

Shalini Santhakumaran
Imperial College London, School of Public Health

A dissertation submitted in partial fulfilment of the requirements of the degree of Doctor of Philosophy

Abstract

Monitoring clinical outcomes across healthcare providers is increasingly important in the UK National Health Service. Neonatal care is no exception, but care is *centralised*, so is delivered via co-ordinated networks of neonatal units (NNUs), with sicker infants treated in larger centres. This results in frequent transfers, but it is unclear how to attribute outcomes of transferred infants. Hierarchical regression is recommended for assessing performance of healthcare providers, but existing studies have either excluded transferred patients or assigned them to a single provider. In this thesis, hierarchical Bayesian multiple membership (MM) models are used to evaluate NNU mortality of very preterm infants, attributing outcomes to all NNUs providing care.

Data for all singleton infants born 2011-2013 below 32 weeks gestation and admitted to neonatal care in England are obtained from the National Neonatal Research Database. Using established methods, a series of Bayesian hierarchical models with two (infants within NNUs) and three levels (infants within NNUs within networks) are developed for non-transferred infants. This approach is extended to include transferred infants using MM models. A variety of weightings, some specified using Beta distributions, are used to allocate outcomes of transferred infants. In contrast to other applications of MM models, results differ across weightings due to transfer patterns. The recommendation is that transferred patients are allocated equally to all providers, regardless of duration or intensity of care, accompanied by sensitivity analyses using alternative weights.

Methodologically, this thesis demonstrates a statistically principled way of accounting for transfers when evaluating provider-specific outcomes, and presents a new application of MM models with novel weightings. Clinically, the variation attributable to providers is low, and for each NNU an estimate of risk-adjusted mortality compared with similar NNUs is obtained. Practical implications extend beyond neonatal medicine as centralisation and electronic patient data become integral to improving healthcare.

Declarations

I declare that all the work presented in this thesis is my own unless otherwise indicated. Any work that is not my own is appropriately referenced, including work done in collaboration with others. Co-authored outputs relating to the thesis are given in appendix G.

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.

Acknowledgements

I would like to thank my supervisors Deborah Ashby and Neena Modi for their wise words and encouragement during this PhD and in other aspects of my work. Thank you to my colleagues, both in the Section of Neonatal Medicine and at the Imperial Clinical Trials Unit, who have all been very supportive and helpful. In particular I would like to thank Chris Gale for reading drafts, Cheryl Battersby for listening, Brad Manktelow for useful discussions early on, and Eugene Statnikov for motivation and cakes.

I would not have completed this thesis were it not for the support of my family and friends. Thank you especially to Mum, Dad, Hilary, James and Elizabeth for keeping Malathi happy and me fed and watered for many weekends.

Most of all I would like to thank Sam. For removing every possible obstacle, for proof-reading, for gourmet dinners, and for never doubting me.

This thesis is dedicated to Malathi.

Abbreviations

AIC Akaike Information Criterion

BAPM British Association of Perinatal Medicine

CI Confidence interval

CrI Credible interval

DH Department of Health

DIC Deviance Information Criterion

HDC High dependency care

IMD Index of multiple deprivation

IC Intensive care

IQR Interquartile range

LNU Local neonatal unit

LOS Length of stay

LSOA Lower super output area

MfN Medicines for Neonates

MCMC Markov Chain Monte Carlo

MM Multiple membership

MSE Mean square error

NDAU Neonatal Data Analysis Unit

NDS Neonatal Dataset

NICU Neonatal intensive care unit

NNAP National Neonatal Audit Program

NNRD National Neonatal Research Database

NNU Neonatal unit

ONS Office of National Statistics

OR Odds ratio

RCS Restricted cubic spline

SE Standard error

SES Socioeconomic status

SGA Small for gestational age

SMR Standardised Mortality Ratio

SC Special care

SCBU Special care baby unit

SD Standard deviation

Contents

1 Introduction **22**

1.1 Preterm birth 23

1.1.1 Definitions, incidence and trends 23

1.1.2 Risk factors for preterm birth 24

1.1.3 Consequences of preterm birth 25

1.2 Neonatal care 28

1.2.1 Managed clinical networks 29

1.2.2 Neonatal transfers 31

1.3 Provider profiling 33

1.3.1 Challenges in provider profiling 34

1.3.2 Hierarchical models for provider profiling 36

1.4 Aims of the thesis 36

1.5 Structure of the thesis 37

2 Data **39**

2.1 The National Neonatal Research Database 39

2.1.1 Development of the NNRD 39

2.1.2 Contents of the NNRD 40

2.1.3 NNRD quality assurance 41

2.2 Selection of data for the thesis 43
2.2.1 Admissions to neonatal care in England 43
2.2.2 Birth weight cohorts 43
2.2.3 Gestational age criteria 44
2.2.4 Congenital malformations 45
2.2.5 Multiple births 46
2.2.6 Defining mortality 47
2.2.7 Missing mortality data 48
2.3 Descriptive analysis 51
2.3.1 Infant characteristics 51
2.3.2 Structure of neonatal care 53
2.3.3 Transfers 54
2.3.4 Mortality 59
2.4 Discussion 62
2.4.1 Summary 62
2.4.2 Data source 62
2.4.3 Selection of the cohort 63
2.4.4 Infant characteristics 65
2.4.5 Structure of neonatal care and transfers 65

3 Risk factors for mortality of preterm infants 67

3.1 Background 67
3.1.1 Risk adjustment for provider profiling 68
3.2 Risk adjustment for the thesis 72
3.2.1 Clinical risk factors 72

3.2.2 Sociodemographic factors 72
3.2.3 Adjusting for transfer status 74
3.2.4 Provider-level variables 75
3.3 Selection of infant risk factors 77
3.3.1 Literature review 77
3.3.2 Description of risk factors 78
3.4 Statistical modelling of infant risk factors 80
3.4.1 Methods 80
3.4.2 Results 85
3.5 Discussion 98
3.5.1 Summary 98
3.5.2 Selection of risk factors 98
3.5.3 Statistical methods 99
3.5.4 NNRD data and associations with mortality 100
3.5.5 Model performance and validation 101

4 Review of hierarchical models and provider profiling methods 103

4.1 Hierarchical models 103
4.1.1 Definition 103
4.1.2 Model estimation 106
4.1.3 Extensions to the basic model 107
4.1.4 Hierarchical models for provider profiling 109
4.2 Literature review of transfers 112
4.2.1 Search strategy 113
4.2.2 Results 114

4.3 Multiple membership models 121
4.3.1 Definition 121
4.3.2 Existing literature on multiple membership models 123
4.4 Discussion 126
4.4.1 Summary 126
4.4.2 Hierachical models 126
4.4.3 Transfers 129
4.4.4 Multiple membership models 131

5 Methods for the thesis **134**

5.1 Bayesian hierarchical models 134
5.1.1 Choice of priors 136
5.1.2 Estimation of Bayesian models 137
5.1.3 Convergence and estimation 137
5.1.4 Model checking and model comparison 138
5.2 Hierarchical analysis of non-transferred infants 140
5.2.1 Aim 140
5.2.2 Data 140
5.2.3 Statistical models 140
5.2.4 Prior distributions 144
5.2.5 Model estimation and checking 145
5.2.6 Presenting results 146
5.3 Multiple membership analysis using simulated data 147
5.3.1 Aim 147
5.3.2 Weighting schemes 147

5.3.3 Data . . . 149
5.3.4 Statistical models and estimation . . . 150
5.3.5 Presenting results . . . 150
5.4 Multiple membership analysis using neonatal data . . . 151
5.4.1 Aim . . . 151
5.4.2 Weighting schemes . . . 151
5.4.3 Data . . . 152
5.4.4 Statistical models and estimation . . . 153
5.4.5 Presenting results . . . 154
5.5 Summary . . . 154

6 Results I: Non-transferred infants 155

6.1 Methods recap . . . 155
6.2 Data . . . 156
6.3 Model checking . . . 158
6.3.1 Convergence . . . 158
6.3.2 Choice of random effect distribution and priors . . . 158
6.3.3 Posterior predictive checks . . . 158
6.3.4 Model fit . . . 163
6.4 Infant and provider characteristics . . . 163
6.5 NNU and network effects . . . 166
6.5.1 Distribution of NNU effects . . . 166
6.5.2 Standard deviation of NNU and network effects . . . 167
6.5.3 Comparison of NNU effects across models . . . 169
6.6 Discussion . . . 176

6.6.1 Summary . . . 176
6.6.2 Model checking . . . 176
6.6.3 Infant and provider characteristics . . . 177
6.6.4 NNU and network effects . . . 179
6.6.5 Clinical summary . . . 181
6.6.6 Implications for the thesis . . . 181

7 Results II: Including transferred infants 183

7.1 Methods recap for simulation study . . . 183
7.2 Results from simulation study . . . 185
7.3 Methods recap for analysis of neonatal data . . . 187
7.4 Results for analysis of neonatal data . . . 188
7.4.1 Data . . . 188
7.4.2 Weight distribution . . . 188
7.4.3 Model checking . . . 189
7.4.4 Infant and provider characteristics . . . 190
7.4.5 NNU and network effects . . . 193
7.5 Discussion . . . 204
7.5.1 Summary . . . 204
7.5.2 Simulation study . . . 205
7.5.3 Analysis of neonatal data . . . 206
7.5.4 Conclusion . . . 208

8 Case study 211

8.1 Aim . . . 211
8.2 Methods . . . 211

8.2.1 Volume . . . 212
8.2.2 Population . . . 212
8.2.3 Mortality . . . 212
8.3 Results . . . 213
8.4 Discussion . . . 214

9 Conclusion **222**
9.1 Research question . . . 222
9.2 Thesis summary . . . 223
9.3 Contributions of the thesis . . . 225
9.4 Discussion . . . 226
9.4.1 Neonatal care . . . 226
9.4.2 Provider profiling . . . 227
9.4.3 Methodological . . . 229
9.5 Further work . . . 230
9.5.1 Missing data . . . 230
9.5.2 Multiple births . . . 231
9.5.3 Other applications . . . 233
9.6 Conclusion . . . 234

A List of neonatal units included in the thesis **262**

B Search results for review of infant risk factors **268**

C Search results for review of patient transfers **274**

D OpenBUGS model code **279**

E Additional results for chapter 6 **286**

E.1 Convergence statistics 286

E.2 Sensitivity analysis with alternative assumptions for NNU effects 293

E.2.1 Normal distribution for NNU effects 293

E.2.2 Caterpillar plots for Normal NNU effects 298

E.2.3 U(0,100) prior for the standard deviation of NNU effects 303

F Additional results for chapter 7 **307**

F.1 Convergence statistics 307

F.2 NNU effects ordered by proportion transferred 310

G List of outputs relating to the thesis **325**

List of Figures

1.1 Livebirths by gestation, England and Wales 2007-2013 24

1.2 Neonatal deaths by gestation, England and Wales 2007-2013 26

1.3 Maps of neonatal units and networks in England 32

2.1 Distribution of gestational age by NNU level 54

2.2 Category of care by NNU level and postnatal age 55

2.3 Transfers by gestational age 57

2.4 Direction of neonatal transfers by level of the referring NNU, singleton infants born below 32 weeks gestation 2011-2013 58

2.5 Timing of neonatal transfers, grouped by direction of transfer for singleton infants born below 32 weeks gestation 2011-2013 59

2.6 Cross-network transfers 60

3.1 Palliative care example plot from the HSCIC 71

3.2 Non linearity of gestational age and birth weight with log odds of death 87

3.3 Predicted probabilities from logistic regression models 88

3.4 Standardised residuals from regression models 93

3.5 Leverage from quadratic birthweight model 94

3.6 Leverage from spline birth weight model 95

3.7 Survival probabilities by gestational age and birth weight 96

4.1 Diagrams illustrating hierarchical data structures 104

4.2 Flowchart illustrating the literature review 114

4.3 Diagrams illustrating multiple membership data structures 122

5.1 Diagram of Beta distribution weights 148

6.1 Posterior predictions by network, model 1 159

6.2 Posterior predictions by network, model 4 160

6.3 Posterior predictions for 24 NNUs, model 1 161

6.4 Posterior predictions by NNU level, models 4 and 5 162

6.5 Q-Q plots for NNU effects, normal and t-4 distributions 166

6.6 Posterior distribution for between provider SD 168

6.7 Correlation of NNU effects 170

6.8 Caterpillar plot for SCBUs 171

6.9 Caterpillar plot for LNU 172

6.10 Caterpillar plot for NICU 173

6.11 Caterpillar plot for networks 174

7.1 Posterior distribution of between-provider SD, NNUs and networks . . . 194

7.2 Posterior distribution of between-NNU SD 195

7.3 Posterior distribution of between-network SD 195

7.4 Correlation of NNU effects across multiple membership weights 198

7.5 Caterpillar plot for SCBU effects 199

7.6 Caterpillar plot for LNU effects 200

7.7 Caterpillar plot for NICU effects 201

7.8 Caterpillar plot for networks effects 202

7.9 Caterpillar plot for LNU effects, ordered by the proportion of infants transferred out within 24 hours 203

8.1 Number of infants admitted by NNU, all NNUs 216

8.2 Number of care days by NNU, all NNUs 216

8.3 Number of IC days by NNU, all NNUs 216

8.4 Number of infants admitted by NNU, NICUs only 217

8.5 Number of infants admitted by NNU, network Y only 217

8.6 Number of care days by NNU, NICUs only 217

8.7 Number of care days by NNU, network Y only 217

8.8 Number of IC days by NNU, NICUs only 217

8.9 Number of IC days by NNU, network Y only 217

8.10 Source of admission for infants admitted to NNU X 217

8.11 Outcome or transfer destination for infants discharged from NNU X . . 217

8.12 Gestational age distribution for NNU X compared with all NNUs 218

8.13 Gestational age distribution for NNU X compared with NICUs only . . 218

8.14 Birth weight distribution for NNU X compared with all NNUs 218

8.15 Birth weight distribution for NNU X compared with NICUs only 218

8.16 Estimated risk of death for NNU X compared with all NNUs 218

8.17 Estimated risk of death for NNU X compared with NICUs only 218

8.18 IMD distribution for NNU X compared with all NNUs 219

8.19 IMD distribution for NNU X compared with NICUs only 219

8.20 Maternal ethnicity distribution for NNU X compared with all NNUs . . 219

8.21 Maternal ethnicity distribution for NNU X compared with NICUs only . 219

8.22 Maternal age distribution for NNU X compared with all NNUs 219

8.23 Maternal age distribution for NNU X compared with NICUs only 219

8.24 Crude mortality for non-transferred infants by NNU, all NNUs 220

8.25 Crude mortality for all infants by NNU, all NNUs 220

8.26 Crude mortality for non-transferred infants by NNU, NICUs only 220

8.27 Crude mortality for all infants by NNU, NICUs only 220

8.28 Adjusted OR for non-transferred infants by NNU, all NNUs 221

8.29 Adjusted OR for non-transferred infants by NNU, NICUs only 221

8.30 Adjusted OR from equal weight MM model by NNU, all NNUs 221

8.31 Adjusted OR from equal weight MM model by NNU, NICUs only 221

8.32 Adjusted OR from weighted LOS MM model by NNU, all NNUs 221

8.33 Adjusted OR from weighted LOS MM model by NNU, NICUs only . . . 221

E.1 Posterior distribution for between provider SD 296

E.2 Correlation of NNU effects across models, Normal distribution for NNU effects with U(0,5) prior for σ_u 297

E.3 Caterpillar plot for SCBU, showing smallest and largest effects across models, Normal NNU effects 299

E.4 Caterpillar plot for LNU, showing smallest and largest effects across models, Normal NNU effects 300

E.5 Caterpillar plot for NICU, showing smallest and largest effects across models, Normal NNU effects 301

E.6 Caterpillar plot for networks, showing smallest and largest effects across models, Normal NNU effects 302

E.7 Posterior distribution for between provider SD 305

E.8 Correlation of NNU effects across models, Normal distribution for NNU effects with U(0,100) prior for σ_u 306

F.1 Caterpillar plots for SCBU effects, ordered by proportion transferred . . 311

F.2 Caterpillar plots for LNU effects, ordered by proportion transferred . . . 312

F.3 Caterpillar plots for NICU effects, ordered by proportion transferred . . 313

F.4 Caterpillar plots for SCBU effects, ordered by proportion transferred in 314

F.5 Caterpillar plots for LNU effects, ordered by proportion transferred in . 315

F.6 Caterpillar plots for NICU effects, ordered by proportion transferred in . 316

F.7 Caterpillar plots for SCBU effects, ordered by proportion transferred out 317

F.8 Caterpillar plots for LNU effects, ordered by proportion transferred out 318

F.9 Caterpillar plots for NICU effects, ordered by proportion transferred out 319

F.10 Caterpillar plots for SCBU effects, ordered by proportion transferred in within 24 hours of birth 320

F.11 Caterpillar plots for LNU effects, ordered by proportion transferred in within 24 hours of birth 321

F.12 Caterpillar plots for NICU effects, ordered by proportion transferred in within 24 hours of birth 322

F.13 Caterpillar plots for SCBU effects, ordered by proportion transferred out within 24 hours of birth 323

F.14 Caterpillar plots for NICU effects, ordered by proportion transferred out within 24 hours of birth 324

List of Tables

2.1 Comparison of infants with and without missing death data in the NNRD 50

2.2 Description of cohort for the thesis by transfer status 52

2.3 Summary of admissions by level of NNU 53

2.4 Summary of type of care by level of NNU 53

2.5 Comparison of reasons for transfer with reason for admission 56

2.6 Mortality across levels of NNU 61

3.1 Distribution of infant risk factors 86

3.2 Akaike Information Criteria to compare number of knots 87

3.3 Comparison of gestational age-birth weight interaction models; birth weight modelled as a 5-knot spline 90

3.4 Comparison of gestational age-birth weight interaction models; birth weight modelled as quadratic 91

3.5 Comparison of odds ratios for final model and sensitivity analysis 92

3.6 Comparison of apparent and optimism-corrected performance statistics . 97

3.7 Comaprison of NNRD model with existing prediction models 97

4.1 Results from systematic review 113

4.2 Results from systematic review *(continues to table 4.3)* 115

4.3 Results from systematic review *(continued from table 4.2)* 116

5.1 Weightings for multiple membership models 152

6.1 Description of cohort for analysis of non-transferred infants 157

6.2 Model performance, t-4 NNU effects U(0,5) prior 163

6.3 [Odds ratios for infant and provider characteristics, t-4 NNU effects U(0,5) prior 165

6.4 Average SD for NNU effects, t-4 distribution for NNU effects, U(0,5) prior 169

7.1 Weightings for multiple membership models 184

7.2 Results from multiple membership simulation study 186

7.3 Description of cohort for analysis of transferred infants 188

7.4 Distribution of multiple membership weight across NNU levels 190

7.5 Distribution of multiple membership weight across episodes, transferred infants only 191

7.6 Model performance 191

7.7 Odds ratios for infant and provider characteristics, MM models 192

7.8 Average SD for NNU effects,MM models 197

A.1 Neonatal units contributing data to the National Neonatal Research Database and included in the thesis *(continues to table A.2)* 263

A.2 Neonatal units contributing data to the National Neonatal Research Database and included in the thesis (*continued from table A.1*) 264

A.3 Neonatal units contributing data to the National Neonatal Research Database and included in the thesis (*continued from table A.2*) 265

A.4 Neonatal units contributing data to the National Neonatal Research Database and included in the thesis (*continued from table A.3*) 266

A.5 Neonatal units contributing data to the National Neonatal Research Database and included in the thesis (*continued from table A.4*) 267

B.1 Results from updating the search of Medlock et al. (2011), continues to table B.2 269

B.2 Results from updating the search of Medlock et al. (2011), continued from table B.1 270

E.1 Monte Carlo standard errors (log scale), t-4 NNU effects U(0,5) prior . . 287

E.2 Estimated sample size, t-4 NNU effects U(0,5) prior 288

E.3 Monte Carlo standard errors (log scale), Normal NNU effects U(0,5) prior 289

E.4 Estimated sample size, Normal NNU effects U(0,5) prior 290

E.5 Monte Carlo standard errors (log scale), Normal NNU effects U(0,100) prior 291

E.6 Estimated sample size, Normal NNU effects U(0,100) prior 292

E.7 Odds ratios for infant and provider characteristics, Normal NNU effects U(0,5) prior 294

E.8 Model performance, Normal NNU effects U(0,5) prior 295

E.9 Model performance, Normal NNU effects U(0,100) prior 303

E.10 Posterior mean odds ratios and 95% credible intervals for infant and provider characteristics, Normal NNU effects U(0,5) prior 304

F.1 Monte Carlo standard errors (log scale) for multiple membership models 308

F.2 Estimated sample size for multiple membership models 309

Chapter 1

Introduction

Whilst most babies are born healthy, some are born too early or with a condition requiring specialist medical attention in a neonatal unit (NNU[1]). In the UK, neonatal care is centralised, so intensive care is not available at every site and infants are frequently transferred. This poses challenges regarding assessment of clinical outcomes across individual NNUs as it is not clear to how to attribute outcomes of transferred infants. Another aspect of centralised care is that neonatal units do not work independently of each other, but rather in co-ordinated groups known as *neonatal networks*. Thus there is a three level structure to the data: infants within NNUs within networks. Hierarchical models are a class of statistical model that is typically used for such structured data, and also for evaluating outcomes across healthcare providers. The aim of this thesis is to address the statistical issues posed by centralisation within the framework of hierarchical models in the context of measuring mortality of preterm infants across neonatal units in England, using data extracted from electronic patient records held in the National Neonatal Research Database (NNRD).

In this chapter I present some context and motivation for the research questions of this thesis. I give a background to preterm birth, describe the structure of neonatal care in England and introduce key concepts in provider profiling, before describing the research question in more detail and giving an outline for the rest of the thesis.

[1]The abbreviation used here is the convention in neonatal care

1.1 Preterm birth

1.1.1 Definitions, incidence and trends

Preterm birth is defined as birth at less than 37^{+0} weeks$^{+\text{days}}$ gestation, with full term birth being 37^{+0} to 42^{+0} (40^{+0} weeks$^{+\text{days}}$ being the estimated due date). Preterm infants are often categorised as extremely preterm (less than 28^{+0} weeks$^{+\text{days}}$), very preterm (28^{+0} to 31^{+6}), moderately preterm (32^{+0} to 33^{+6}) and late preterm (34^{+0} to 36^{+6})[2](Blencowe et al., 2012). Preterm birth can occur spontaneously, or because risks to the fetus and/or the mother prompt a clinical decision to deliver early (*medically indicated preterm birth*).

International estimates of the preterm birth rate range from 5% to 18% of livebirths, with higher rates in low-income countries (Blencowe et al., 2012). Within Europe, the preterm birth rate varies from 5% to 10% of livebirths (Gissler and Mohangoo, 2010). In England and Wales between 2007 and 2013, 7.2% of livebirths of known gestational age were preterm (Office of National Statistics, 2014). Figure 1.1 shows the distribution of births by gestational age based on Office of National Statistics (ONS) data on livebirths in England and Wales. As the numbers born preterm are low, particularly less than 32 weeks gestation, I have shown the distribution over all births, all preterm births and all very preterm birth separately.

Globally, there has been an increase in the preterm birth rate in recent decades (Blencowe et al., 2012), though trends are not uniform (Zeitlin et al., 2013). In high-income countries, increased use of fertility treatments and advanced maternal age have led to higher multiple pregnancy rates, and multiple births are five to ten times more likely to be delivered preterm compared with singletons (Blondel et al., 2002; Zeitlin et al., 2013). This may be a key driver for the increasing preterm birth rate in some areas: in half the countries in the Euro-Peristat project the preterm birth rate amongst singletons remained stable or declined despite overall increases (Zeitlin et al., 2013). Some of the increase in preterm birth may be driven by higher rates of medically indicated late and moderate preterm birth, due to improved risk assessment and timely diagnosis of disorders in both mother and baby (Goldenberg et al., 2008; Shapiro-Mendoza and Lackritz, 2012). In some countries there is evidence that the preterm birth rate is slowing down and even declining, albeit slightly (for example the USA (Child Trends Databank, 2015), Finland (Jakobsson et al., 2008) and the Netherlands (Schaaf et al., 2011)). Nevertheless, increases in overall fertility mean that the absolute number of preterm births continues to increase (Department of Health, 2009; Goldenberg et al., 2008; Saigal and Doyle, 2008).

[2]Henceforth I will express the units as *weeks* only, with no superscript being shorthand for zero days (so 31^{+6} weeks$^{+\text{days}}$ will be written as 31^{+6} weeks and 31^{+0} weeks$^{+\text{days}}$ will be written as 31 weeks)

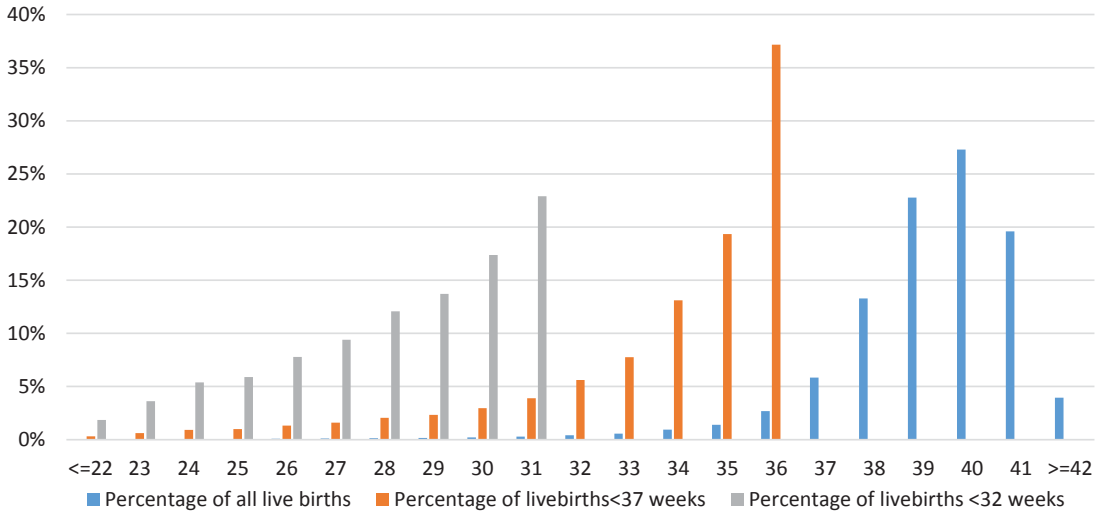


Figure 1.1: Percentage of all livebirths (blue), all preterm livebirths (orange) and all very preterm livebirths (grey) born at each week of gestation, England and Wales 2007-2013. Data from the Office for National Statistics licensed under the Open Government Licence v.3.0.

1.1.2 Risk factors for preterm birth

The aetiology of spontaneous preterm birth is not fully understood, but it is thought to be a syndrome induced by infection or inflammation, stress, and other immunologically mediated processes (Goldenberg et al., 2008). While there are known clinical risk factors such as family history, maternal hypertension and short interpregnancy interval (Blencowe et al., 2012; Goldenberg et al., 2008) the causal mechanisms linking these factors to preterm birth remain somewhat unclear, and it is therefore difficult to attribute preterm birth to the various potential causes. A number of sociodemographic factors are associated with preterm birth such as stress, smoking, maternal age and ethnicity. These are also linked to deprivation, leading to higher rates of preterm delivery in more deprived areas (Gray and McCormick, 2009; Kramer et al., 2001).

Other risk factors include medical indications for early delivery such as pre-eclampsia and HELLP syndrome (maternal conditions related to hypertension), growth restriction in the fetus and placental problems. The definitions of "indicated" and "spontaneous" are not consistently applied, nor are the medical indications for early delivery always clear, rendering classification difficult (Goldenberg and Gravett, 2012).

As mentioned previously, multiple birth can lead to preterm delivery, both spontaneous and medically indicated. In England in 2013, 24% of infants born preterm were

from a multiple pregnancy, in contrast with 1.5% of term births (Office of National Statistics, 2014). There is some indication that singletons and multiples have a different distribution of gestational age, and possibly a different "optimal" gestation at around 38 weeks (Buekens and Wilcox, 1993; Kiely, 1998; Minakami and Sato, 1996). Reasons for indicated early delivery may differ between multiples and singletons (Gardner et al., 1995). Therefore the mechanisms resulting in early delivery of multiple births are not the same as those for singletons.

1.1.3 Consequences of preterm birth

When infants are born too early their organs, in particular the brain and lungs, are insufficiently developed and not ready for life outside the womb. This has a number of short and long term health consequences.

Mortality

Globally, preterm birth is responsible for over a third of neonatal deaths (defined as death within 28 days of birth). The influence of preterm birth extends into childhood, and it is the second most common cause of death amongst children under 5 (World Health Organization, 2012). Infants born preterm have a much higher risk of mortality than their term-born counterparts, and the causes of death are quite different, with respiratory problems and ventricular haemorrhage being the main factors in preterms (Sidebotham et al., 2014). The mortality difference decreases with increasing gestational age, though an additional risk persists even for late preterms. Figure 1.2 shows neonatal deaths by gestational age based on ONS data for England and Wales 2007 to 2013, both as numbers and as a percentage of all live births at that gestational age week. Although figure 1.1 showed that birth at or below 32 weeks gestation is rare, hovering at just above 1% for the past few years in England and Wales, this gestational age group represent the majority of neonatal deaths with a mortality rate of 12.1%.

Mortality of preterm infants varies widely globally: the risk of neonatal death is 12 times higher for an infant born preterm in Africa compared with Europe (World Health Organization, 2012). Even within high income countries there is considerable variation. MacDorman et al. (2014) compared data from the US and Europe using three sources (US Vital Statistics, the Organisation for Economic Co-operation and Development database and the European Perinatal Health Report) and found that infant mortality (death within 1 year) varied for births between 24-27 weeks (16.5% in Sweden to 42.9% in Poland) and 28-31 weeks (3.4% in Austria to 9.9% in Poland). However, variations in stillbirth definition and resuscitation of infants at the limits of viability, amongst other

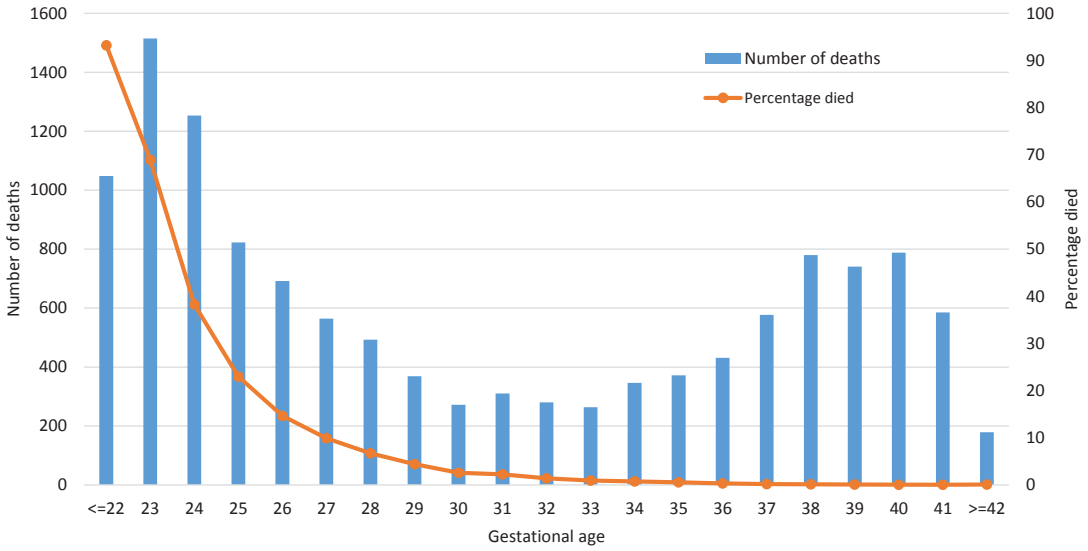


Figure 1.2: Number (blue) and percentage (orange) of neonatal deaths by week of gestation, for all livebirths in England and Wales 2007-2013. Adapted from data from the Office for National Statistics licensed under the Open Government Licence v.3.0.

factors, limit the utility of such comparisons (Draper et al., 2009; Joseph et al., 2012).

In previous decades there have been a number of significant medical advances leading to substantial reductions in mortality of preterm infants. In 1975 half of all liveborn preterm infants with a birth weight less than 1500g died, but by 1985 this had reduced to a quarter, and to a sixth by 1995 (National Audit Office, 2007). Lung immaturity is one of the most severe consequences of preterm birth and is a major contributing factor to mortality of preterm infants. Surfactant is a substance which lines the alveoli (air sacs) where oxygen is transferred into the blood and carbon dioxide is taken out. Without it, there is too much surface tension, and with the pressure of exhalation the alveoli collapse in on themselves, making inhalation difficult. Surfactant occurs naturally in the lungs, but the levels build up in the last trimester of pregnancy so infants born preterm may not have sufficient quantities, resulting in respiratory distress syndrome. Surfactant replacement, along with the provision of antenatal steroids to the mother to accelerate fetal lung development, resulted in substantial reductions in mortality at the end of the 20th century (Halliday, 2008).

More recent trends are not as dramatic as those seen in the 1980s and 1990s, but there is evidence that mortality of preterm infants continues to decline in the UK, particularly for the earliest gestations. The EPICure studies, which examined short and long term outcomes of extremely preterm infants in the UK, found a reduction in death before discharge from 60% in 1995 to 47% in 2006 for infants born between 22^{+0} and

25^{+6} weeks gestation (Costeloe et al., 2012). Improvements have also been observed in population data from the US (Horbar et al., 2012; Patel et al., 2015), France (Ancel et al., 2015) and Australia (Bolisetty et al., 2015). There have been a number of strategies which may have contributed to further reductions in preterm mortality, including centralisation of services (Costeloe et al., 2012). Centralisation involves shifting services away from smaller providers to fewer, larger centres. The rationale for centralisation of neonatal care is that infants with the most complex needs have better survival prospects at experienced, high-volume centres (Cifuentes et al., 2002; Lasswell et al., 2010; Phibbs et al., 1996; Watson et al., 2014). It does not necessarily follow that adopting a centralised service will result in improved outcomes for all, one reason being that the increased burden on the larger centres may outweigh some of the benefit. Although there have been studies examining changes in survival before and after centralisation (Binder et al., 2011; Neto, 2006) it is difficult to attribute improvements to centralisation policies as there are likely to have been other concurrent changes contributing to the reductions (Gale et al., 2012b). The centralised structure of neonatal services in the UK will be discussed further in section 1.2.

Other consequences

With survival of preterm infants improving, the sequelae for survivors are becoming increasingly evident. Preterm birth can lead to neurodevelopmental impairment with consequences for learning, behaviour and mental health in later life. Cerebral palsy is often seen as a marker of intrapartum and neonatal care (Murphy et al., 1997; Saigal and Doyle, 2008) and rates are higher in children born preterm; estimates from a meta-analysis by Himpens et al. (2008) were 14.6% for extremely preterm and 6.2% for very preterm, compared with 0.1% in term-born children. Cognitive and sensory impairments are also higher following preterm birth (Saigal and Doyle, 2008).

While problems associated with lung maturation have been alleviated by steroids, surfactant, and assisted ventilation, preterm infants remain susceptible to respiratory problems. Bronchopulmonary dysplasia (BPD), which is associated with prolonged ventilator use, affects between 10% and 25% of very preterm infants and is often defined by the requirement for oxygen therapy, usually at 36 weeks postmenstrual age (Gortner et al., 2011).

Necrotising enterocolitis (NEC) is a severe disease affecting the gastrointestinal tract which primarily affects extremely preterm infants. In preliminary findings from the UK Neonatal Collaborative NEC study, Battersby et al. (2014a) found 8% of infants born 23^{+0} to 25^{+6} weeks developed severe NEC requiring surgery, compared with 0.8% of infants born 29^{+0} to 31^{+6} weeks with 40% of these infants dying before discharge from

neonatal care. The rarity of the condition coupled with the lack of an agreed clinical definition makes establishing preventative strategies through research difficult.

As our understanding of the long-term consequences for preterm survivors has increased, more attention has been paid to preventing preterm birth. The wide variety of potential reasons for preterm birth combined with the lack of clinical understanding of the causal mechanisms has made it a difficult problem to tackle. While preventative strategies are in use, such as tocolytic drugs to delay the onset of labour, these are largely unsuccessful in improving clinical outcomes (Alfirevic, 2012; Zeitlin et al., 2013). Until effective interventions emerge, preterm birth is a public health issue which shows no sign of abating.

1.2 Neonatal care

Neonatal care refers to specialist treatment in hospital that a newborn may require beyond routine postnatal care. Whilst most babies are born healthy, some are born too early or with a condition requiring specialist medical attention in a neonatal unit. In 2014 there were over 86,000 admissions to neonatal specialist care in England representing 1 in 8 live births (Royal College of Paediatrics and Child Health, 2015). Preterm birth and its associated sequelae account for around half of these admissions; other reasons include low birth weight, congenital malformations and complications during pregnancy and delivery. Neonatal medicine is a growing speciality. Increases in both overall fertility and the rate of preterm birth have driven a rise in the number of babies requiring admission to a neonatal unit (Department of Health, 2009; Goldenberg et al., 2008; Saigal and Doyle, 2008). As discussed in section 1.1, improvements in medical care have led to better chances of survival for babies born preterm. In particular the proportion of infants born at the lower end of the gestational age range has increased dramatically, along with their survival rates (Costeloe et al., 2012). This combination of increased preterm admissions coupled with improved survival rates due to medical and technological advances has led to escalating pressure on neonatal services (Department of Health, 2009; Goldenberg et al., 2008; National Audit Office, 2007; Tucker and McGuire, 2004).

One of the challenges of managing neonatal services is the heterogeneous needs of the patient population. The type of care required can be grouped into three categories: special care (SC), high dependency care (HDC) and intensive care (IC) using definitions developed by British Association of Perinatal Medicine (BAPM)(2011)[3]. The category of care (which is an integral part of determining resource use and funding for each NNU,

[3]A new version of these categories is to be published this year (2016) but is not yet in use, and will not be retrospectively applied.

see Department of Health (2009)) is determined daily by applying exact definitions based on the procedures and treatments required each day. Below are the general principles for the categories of care as given by the BAPM guidelines (2011, pp. 3-4):

Special Care Provided for babies who require additional care delivered by the neonatal service but do not require either Intensive or High Dependency care. Qualifying treatments/conditions include oxygen therapy via a nasal cannula, feeding by nasogastric tube and phototherapy.

High Dependency Care Provided for babies who require highly skilled staff but where the ratio of nurse to patient is less than intensive care. Qualifying treatments/conditions include non-invasive respiratory support, parenteral nutrition and seizures.

Intensive Care Provided for babies who are the most unwell or unstable and have the greatest needs in relation to staff skills and staff to patient ratios. Qualifying treatments/conditions include mechanical ventilation, umbilical lines and any surgery. In the event of death, the day of death is considered an intensive care day.

The staff to patient ratios currently recommended by BAPM are 1:1 for intensive care, 1:2 for high dependency care and 1:4 for special care (Department of Health, 2009). While staffing shortages mean these are not always adhered to (Redshaw and Hamilton, 2005), each NNU should be working towards these standards as part of agreements with service commissioners (Neonatal Critical Care Clinical Reference Group, 2015).

Neonatal intensive care is a particularly high-cost and low-volume speciality, so not all types of care are available at every hospital with a maternity unit. Furthermore, the need for neonatal specialist care is not always known prior to delivery, so postnatal transfer is unavoidable in some cases. The category of care is dependent on treatment received that day so a baby's requirement can change over time, also leading to transfers. I will examine types of transfer in more detail in section 1.2.2. Whatever the reason for transfer, proper organisation of services is critical to ensure babies receive the care they need close to home.

1.2.1 Restructuring of neonatal care into managed clinical networks

In response to the rise in both the size and complexity of care demands, the Department of Health (DH) commissioned a review into the provision of neonatal specialist care in 2001. The working group published the resulting "National Strategy for Improvement" in 2003 (Department of Health, 2003).

The existing pattern of transfers was found to be unsatisfactory, with infants being cared for far from home, so parents had to travel long distances or stay nearby. These problems were exacerbated by limitations in capacity at the higher level NNUs. In addition, there was an incoherence in the pattern of delivery of care, with a lack of national standards to work to. DH recommended that care be provided within *managed clinical networks* with the aim that transfers outside the network would be minimal, and that communication and organisation within networks would improve, resulting in improved transfer patterns. Their conclusions were in line with a number of studies before and since, advocating centralisation of services by demonstrating that preterm or low birth weight infants have better survival prospects at high volume NNUs (Cifuentes et al., 2002; Lasswell et al., 2010; Phibbs et al., 1996; Watson et al., 2014). However rather than simply closing smaller NNUs (as has been the case in some countries, see Phibbs (2012)), DH proposed the networked model of care so as to maintain equity of access and limit the burden of travelling for families of infants in neonatal care, as transfers to more distant units would only be required for the sickest infants.

In 2007 the National Audit Office conducted a review of the reorganisation of neonatal services (National Audit Office, 2007). They examined how well the network structure had been established and whether it had resulted in improvements in how the service functioned. They found varying degrees of success in establishing networks, but reductions in long-distance transfers had been achieved. However there were still improvements to be made, including consistency in transfer arrangements across networks. In response to this DH set up a Neonatal Taskforce, who issued the Neonatal Toolkit (2009), which set forth eight principles to help guide commissioners and practitioners of neonatal care, and provided further guidance on delivery of care under the networked model.

To ensure that all services are available, the networks are comprised of different types of neonatal units providing different types of care. These categories were redefined in the the Neonatal Toolkit as special care baby units [4], local neonatal units, and neonatal intensive care units; however there is crossover between the types of units and the care provided. The units are defined as follows, taken from the Neonatal Toolkit (2009, p. 10):

Special care baby units (SCBUs) These provide special care for their own local population. Depending on arrangements within their network, they may also provide some high dependency services. In addition, SCBUs provide a stabilisation facility for babies who need to be transferred to a NICU for intensive or high dependency care, and they also receive transfers from other units in the network for continuing special

[4]The terminology has now changed to *special care unit* but the definitions for the duration of the data used in this thesis are as given here.

care.

Local neonatal units (LNUs) These provide neonatal care for their own catchment population, except for the sickest babies. They provide all categories of neonatal care, but they transfer babies who require complex or longer-term intensive care to a NICU, as they are not staffed to provide longer-term intensive care. The majority of babies over 27 weeks of gestation will usually receive their full care, including short periods of intensive care, within their LNU. Some networks have agreed variations on this policy, due to local requirements. Some LNUs provide high dependency care and short periods of intensive care for their network population. LNUs may receive transfers from other neonatal services in the network, if these fall within their agreed work pattern.

Neonatal intensive care units (NICUs) These provide the whole range of medical neonatal care for their local population, along with additional care for babies and their families referred from the network. Many NICUs in England are co-located with neonatal surgery services and other specialised services.

In 2013 there were 179 NNUs in England and Wales across 24 networks (Royal College of Paediatrics and Child Health, 2014). These networks have now been further revised but in this thesis I will examine the managed clinical networks as this was the main system in place for the duration of the data analysed (implications of the change for future analysis will be discussion in chapter 9). The NNUs and networks are shown in figure 1.3 and a list is given in appendix A. Note that the boundaries in figure 1.3 are approximated based on mother's postcode as neonatal networks do not in general have strict geographical boundaries.

1.2.2 Neonatal transfers

The change to a networked model of care has brought with it an increase in transfers; indeed transfers and their organisation is seen as key aspect in how networks operate (Redshaw and Hamilton, 2006). Comparison of transfer rates before and after re-organisation, for infants born 27-28 weeks gestation, showed an increase from 25% to 34% (Gale et al., 2012b), but changes have been inconsistent across regions (Cusack et al., 2007; Kempley et al., 2006). National guidance from the NHS Neonatal Critical Care Clinical Reference Group (2015) suggests that care should be provided by a NICU when birth before 27 weeks gestation (28 for multiples) is anticipated, though there may be local agreed variations. If preterm birth is medically indicated before labour commences then delivery is arranged at a hospital with a NICU. Spontaneous deliveries may involve a transfer if the mother presents at a SCBU or LNU, depending on the care required. Transfer of infants requiring specialist care can either occur before birth (*in-*

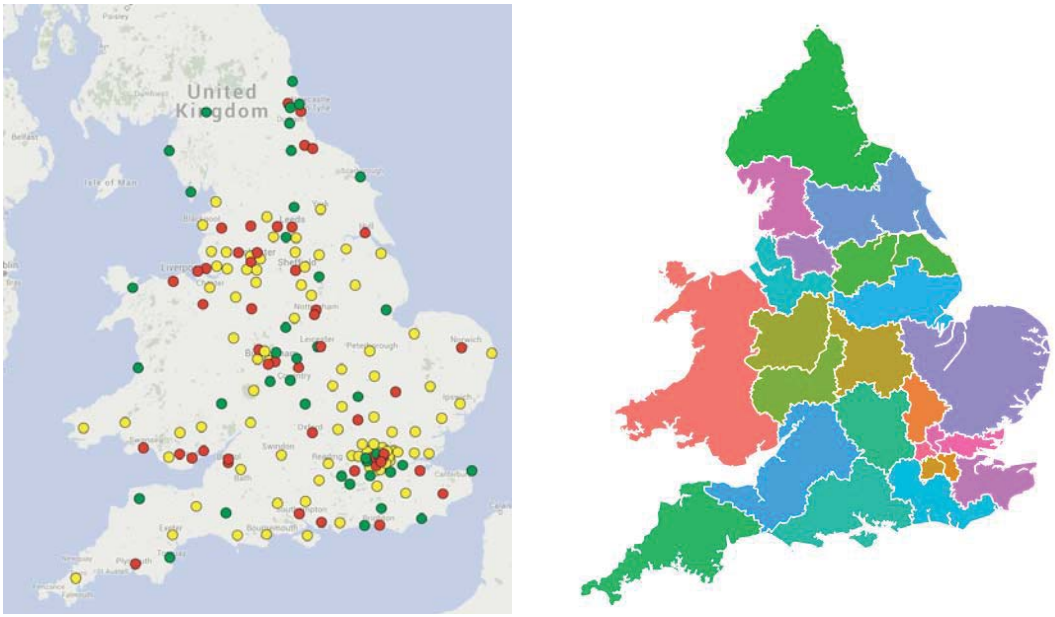


Figure 1.3: a) NNUs in England and Wales with red dots indicating NICUs, yellow for LNUs and green for SCBUs. (Neonatal Data Analysis Unit, 2016), map data © 2016 Google b) neonatal networks, with boundaries estimated based on area of residence of the mother; data generating maps courtesy of the Neonatal Data Analysis Unit and the North West London Collaboration for Leadership in Applied Health Research and Care

utero transfer) or after birth (*postnatal transfer*). In-utero transfer leads to improved infant outcomes (Kollée et al., 1988; Lamont et al., 1983; Shlossman et al., 1997) but is not feasible in all circumstances. If the need for specialist care is not known prior to delivery then postnatal transfer is the only option. Even if the need is known at the onset of labour (as is the case for extremely preterm infants), in-utero transfer may not be possible due to the imminence of birth, the condition of the mother and baby, or logistical difficulties in finding a cot and arranging transport (Gale et al., 2012a; Kempley et al., 2006).

Transfers can occur for several reasons. Although no official classification exists,Gill et al. (2004, p. F221) summarise transfers as follows:

Acute Emergency transfer for more appropriate care

Specialist Non-emergency transfer for a specialist elective service such as surgery

Non-acute Transfer back to the local neonatal unit for step-down care

Capacity Transfer due to lack of space at current unit

The network system was set up to maximise the proportion of high-needs infants being delivered at the appropriate location and to minimise inappropriate transfers (such as

capacity and long-distance transfers) through improved co-ordination between NNUs. There are conflicting reports as to whether these aims have been achieved. Gale et al. (2012b) found that under networked care, 49% of infants born 27-28 weeks were delivered at high-volume NNUs, in contrast to 18% beforehand. Kempley et al. (2006) found that transfer times in the South-East of England had improved but this was in conjunction with the introduction of a regional transfer service. It was accompanied by a decline in in-utero transfers, echoed by Cusack et al. (2007), who in addition found that almost a quarter of transfers were for inappropriate reasons. Neonatal networks are organised in such a way that infants can, in general, receive all the care they require in one network. This is not always the case (17% of all transfers are to a different neonatal network, see Royal College of Paediatrics and Child Health (2015)) but there are legitimate reasons for this. The recommended transfer pathway may incorporate cross-network transfers because the geographical distances are closer than those within the network, or due to the patterns of clinical need in a region. If a mother delivers unexpectedly at a location other than where she was booked to deliver (as is more likely to be the case with spontaneous preterm births compared with term due to being unexpected), the infant may be transferred across networks to receive care closer to home.

1.3 Provider profiling

Provider profiling refers to the estimation, assessment, and comparison of outcomes, process measures and other performance indicators across multiple providers of public services. Over the past few decades these types of analysis have proliferated both in the UK and internationally (Bird et al., 2005; Burgess et al., 2000; Goldstein and Spiegelhalter, 1996; Normand et al., 1997; Spiegelhalter et al., 2012). Increasing importance is being placed on monitoring healthcare outcomes in the UK NHS, as evidenced by the Department of Health (DH) White Paper (Department of Health, 2010). Neonatal care is no exception: the Neonatal Toolkit (2009) recommended that measurement and publication of clinical outcomes was a key element of providing high quality services. Part of the surge in provider profiling, and particularly in placing results in the public domain, is due to a the shift in the way public services operate, with an increased focus on transparency, accountability and choice (Bird et al., 2005). For example, the NHS Choices website allows members of the public to examine data on outcome measures such as mortality, and process measures such as waiting times, for any NHS provider (NHS Choices). Other purposes include examining the performance of individual providers, comparing across different providers, measuring overall variation, establishing suitable performance targets, and identifying aberrant performance. Collecting and analysing suitable data in a way which is appropriate for the purpose can be burdensome (Spiegelhalter et al., 2012). Such analyses are often used by multiple stakeholders (patients

and the general public, clincians, managers and policy-makers) and put to several of the purposes previously described. However, as with any statistical analyses, provider profiling studies are designed to estimate a specific quantity or test a specific hypothesis. Thus a single analysis cannot usually serve multiple purposes adequately (Bird et al., 2005), and the inferences we can draw depend on the assumptions and properties of the methods used (Shahian and Normand, 2015; Shen and Louis, 1998). In chapter 4 I will consider what inferences are feasible within the confines of the data and analysis methods used in this thesis.

1.3.1 Challenges in provider profiling

Provider profiling in healthcare is not a recent innovation. Florence Nightingale, while better known by most for her nursing reforms, has been lauded by the statistical community for her meticulous data collection and novel graphical displays, illustrating variations across hospitals amongst other objectives (Kopf, 1916). As noted by Spiegelhalter (1999), many of the challenges today were documented by Nightingale over 150 years ago.

Case-mix differences

> *"...so many operations of such and such a nature, without reference to age, sex or cause of the operation, followed by so many deaths, without reference to age, sex or complications. Given these elements, divide the one by the other, and you get the mortality. A statistical proceeding such as this can at the best lead to very loose approximations."* Nightingale (1863, p. 171)

In order to make provider profiling analyses meaningful the *signal* (factors we are interested in) needs to be identified separately from the *noise* (Lilford and Pronovost, 2010). Differences in case-mix (the age, sex and cause of the operation in Nightingale's case) represent a major source of noise. All else being equal, hospitals treating sicker patients will have poorer outcomes. In order to draw fair comparisons across providers, we need to control for patient characteristics Iezzoni (2003). We can never truly "level the playing field" as it is rarely possible to know about, let alone accurately measure and control for, all case-mix variables. However we can minimise variation using statistical techniques such as regression. Regression allows us to model a variable of interest (e.g. the probability of death before discharge) as a function of different variables (e.g. a measure of illness severity at admission). This will be discussed further in chapter 3.

Misconception of "excess" deaths

> *"It can convey but a very imperfect idea of the real state of the case. And one thing is quite certain, it can lead to no practical result whatever, either as regards the true causes of the mortality, or how these might be mitigated"* Nightingale (1863, p. 171)

If a provider has a higher mortality than is expected, these are often termed *excess deaths*, and frequently interpreted as therefore being avoidable (Lilford and Pronovost, 2010). We can determine whether there are differences in mortality rates and we can control for some case-mix differences. However we cannot understand (at least not from routinely collected data) all the reasons why a death occurred or whether it could have been prevented. Therefore if a provider has a high risk-adjusted mortality we cannot conclude that this is because they are providing poor-quality care (Goldstein and Spiegelhalter, 1996; Mohammed et al., 2009). This aspect is a source of notoriety of provider profiling studies, with much debate surrounding the reliability of outlier detection, the link between bad results and substandard care, and the utility of naming outliers as a mechanism to drive improvement (Hofer et al., 1999; Lilford and Pronovost, 2010; Marshall et al., 1998; Pitches et al., 2007). Large-scale epidemiological analysis should not be seen as the final word but rather a screening tool which prompts small-scale clinical case review (Jones et al., 2008; Spiegelhalter, 1999), such as the confidential enquiry process in the UK (Manktelow et al., 2016). While potentially subjective, examination of individual cases is generally required for a fuller understanding the circumstances leading up to a death (Sidebotham et al., 2014).

Playing the system

> *"We have known incurable cases discharged from one hospital, to which the deaths ought to have been accounted and received into another hospital, to die there in a day or two after admission thereby lowering the mortality rate of the first at the expense of the second"* Nightingale (1863, p. 171)

Though Nightingale's example may seem extreme, similar actions have been observed in more recent times. In a review of New York cardiac surgery profiling initiative, Normand and Shahian (2007) documented findings from several studies where higher risk patients (or those requiring higher-risk procedures) were referred elsewhere or even refused treatment. More commonly, gaming occurs in the form of data entry, with Green and Wintfeld (1995) finding that 41% of the apparent improvement in mortality following coronary artery bypass graft could be explained by overstating risk factors and the application of different definitions. The choice of risk factors and data sources from which they are obtained can have help reduce the risk of gaming; this will be discussed in chapter 3.

1.3.2 Hierarchical models for provider profiling

Regression models are often used as part of the risk-adjustment process (Iezzoni, 2003). One of the assumptions behind regression is that all observations are independent, conditional on model covariates. This is not the case for provider profiling as patients who are treated by the same provider are likely to have similar outcomes; in fact the influence of a provider on the outcome is the very thing we are interested in. Ignoring the structure of the data by assuming independence leads to standard errors which are underestimated for provider-level variables, and overestimated for patient-level variables (Rabe-Hesketh and Skrondal, 2005). We could include a variable to estimate an independent effect for each provider (called a *fixed effects* model)but this is not feasible when there are many providers, and may mean that small providers with only a few observations have to be excluded. Hierarchical models specifically model the effect of the cluster (in this case the provider) on the outcome, but rather than have each provider effect being independent (equivalent to the fixed effects model) we assume that provider effects are linked with a common distribution. How we specify the structure of the model and what factors we adjust for is driven by the purposes of the model, and hence the inferences we wish to make about providers. I will describe hierarchical modelling in more detail in chapter 4 along with its advantages for the purposes of provider profiling.

1.4 Aims of the thesis

Measurement of clinical outcomes (*provider profiling*) after admission to neonatal care is important for sharing best practice and identifying areas for improvement (Department of Health, 2009). Determining outcomes such as mortality rates for individual neonatal units (NNUs) in the UK is complicated by the structure of neonatal services. There are three elements to this which I shall investigate in this thesis:

Transfers Not all services are provided at all NNUs so infants are often transferred to receive the care they need, maybe multiple times. This raises questions for provider profiling analyses as it is not obvious to which NNU the infant's outcome should be assigned. Infants requiring acute or specialist transfer are likely to be sicker than otherwise similar infants who do not (Delaney-Black et al., 1989; Hohlagschwandtner et al., 2001), so transfer could be considered as a risk factor to adjust for. Although rare, deliberate transfers of sick patients to avoid a death have been documented in other areas (Normand and Shahian, 2007), so the potential for outcome monitoring in neonatal care to influence transfer decisions should be considered.

Type of neonatal unit Each neonatal unit is classed as a special care baby unit, local neonatal unit or neonatal intensive care unit depending on the type of care it provides. It may not be reasonable to compare NNUs across different levels as they are not designed to provide the same care. Infants are likely to be transferred across levels as their clinical needs change, so analysis cannot be conducted separately within levels.

Neonatal networks Neonatal care is delivered via managed clinical networks with each NNU being part of a network. Networks are not simply groups of NNUs, but organisations with their own protocols, standards and care pathways as well as internal procedures for training and clinical governance (Department of Health, 2003; National Audit Office, 2007; Redshaw and Hamilton, 2006), so it seems plausible that we would see correlated outcomes within networks in contrast to providers which do not coordinate to provide care. Infants are already grouped within NNUs, so the organisation of NNUs into networks creates a third level in the data (infants within NNUs within networks). The dependencies between observations need to be correctly accounted for, but we are also interested in measuring outcomes at the network level in addition to the NNU level. Although networks are designed to provide comprehensive care, transfer to another network can happen for a number of reasons. Cross-network transfers add another complication as infants can be treated by multiple networks, not just multiple NNUs.

The aim of this thesis is to establish suitable statistical models to inform NNUs how the expected mortality for infants in their care compares with that of similar infants attending similar NNUs. In particular I investigate assignment of the outcomes of transferred patients, modelling differences across NNU levels and handling of data dependencies at the neonatal network level.

1.5 Structure of the thesis

In chapter 2 I introduce the data source for the thesis, select the cohort for analysis and define the outcome. The data are described with respect to structural aspects of the data that are key to the research questions, namely transfers, NNU levels and neonatal networks. Infant risk factors for mortality are explored in chapter 3, with selection based on review of the literature and consideration of appropriate variables for provider profiling. A risk adjustment model is developed and evaluated, and provider-level variables are considered. In chapter 4 I introduce hierarchical models, describe their use in provider profiling studies and the extent to which standard methods can be applied to the questions of the thesis. I review the literature for existing approaches to the problems presented here, identify gaps in current practice, and introduce multiple membership models as a potential solution. Methods applied in the thesis are presented in

chapter 5. I first consider non-transferred infants only, presenting a series of hierarchical models incorporating NNU and network effects in light of the descriptive analysis in chapter 2 and including the risk-adjustment model developed in chapter 3. A simulation study is described to explore the impact of different multiple membership models in data where the patterns of dependency are known. I then present multiple membership models which allow for transfers, building on the methods for the analysis of non-transferred infants. Results are presented in chapters 6 (non-transferred infants) and 7 (multiple membership models for simulated and neonatal data), and findings summarised in the context of existing literature and the rest of the thesis. Chapter 8 illustrates results for a single NNU as an example of how the findings from this thesis could be presented for clinical purposes. Finally, in chapter 9 I give an overview of the whole thesis, including limitations and further work.

Chapter 2

Data

In section 2.1, I introduce the data source for the thesis, the National Neonatal Research Database (NNRD), including its development, uses and content. Section 2.2 describes the selection of the cohort of infants analysed in this thesis and definition of the outcome of interest. Descriptive analysis of the cohort is provided in section 2.3. I present infant characteristics (section 2.3.1), describe data by neonatal unit level (section 2.3.2), explore transfers (section 2.3.3) and mortality (section 2.3.4). In section 2.4 I summarise and discuss the findings of the chapter with relevance to other literature and implications for the rest of the thesis. The purpose of this chapter is to explore data relating to the structure of services and to provide background for the provider profiling analysis in later chapters.

2.1 The National Neonatal Research Database

2.1.1 Development of the NNRD

As discussed in chapter 1, infants receiving neonatal specialist care frequently undergo transfer as not all services are available at every hospital. Their medical records need to be available at every hospital that cares for them. The Department of Health (DH) review of neonatal services (2003) found that there was a lack of outcome data at a national level and recommended that IT infrastructure be improved to facilitate data collection. These requirements led to the recommendation that the clinical record for infants in neonatal care be stored electronically, in order to be easily available should the infant be transferred, and to facilitate outcome reporting. In 2004, a system to capture data electronically at the point of care was launched in the South-East of England, and has since extended to all NNUs in England, Wales and Scotland, replacing previous local

data collections. The database operates on a web-based platform hosted and managed by Clevermed Ltd. on NHS.net, allowing data to be shared across NNUs. Neonatal units and networks can access data for the infants in their care, enabling local audit and other regional initiatives. Collating the records on a national scale provides the opportunity to use these data for purposes outside day-to-day clinical care, in particular clinical research. In pursuit of this aim, the Neonatal Data Analysis Unit (NDAU) was launched in 2007 with the purpose of supporting neonatal research and promoting data quality. The NDAU is an academic unit which is part of Imperial College London and based at the Chelsea and Westminster NHS Foundation Trust. Data from all NNUs using the system are extracted quarterly by Clevermed and cleaned by NDAU to form the NNRD.

The development of the NNRD from an electronic patient record to a national resource for clinical research was the primary aim of the Medicines for Neonates (MfN) Research programme, funded by the National Institute of Health Resarch (Modi et al., 2016). The programme, completed in March 2015, involved the expansion of the NNRD to cover all NNUs in England, Wales and Scotland, the creation of the Neonatal Data Set as a new NHS information standard (see section 2.1.2) and the assessment of the utility of the NNRD for clinical research in a number of settings. The NNRD has approval as a database from the National Research Ethics Service (10/80803/151). Each NHS Trust contributing data has given Caldicott Guardian approval for data from their Trust to be used for service evaluations by NDAU, including comparisons of mortality performed by NDAU of which this PhD is an extension. The NNRD is used for a wide variety of purposes including evaluation of the restructuring of neonatal care (Gale et al., 2012b) and quality improvement initiatives (Battersby et al., 2014b), development of growth charts for preterm infants (Cole et al., 2013), the National Neonatal Audit Program (NNAP) (Royal College of Paediatrics and Child Health, 2015) and international collaborations (Shah et al., 2014), with integration into data collection for clinical trials in the pipeline (Gale et al., 2016a). The NNRD is the only dataset of its kind in the UK that provides national coverage of neonatal clinical data. Similar databases exist in other countries such as Sweden (EXPRESS Group et al., 2009), the USA (Tyson et al., 2008) and Japan (Kusuda et al., 2012) but these do not cover all admissions to neonatal care nationwide. Furthermore, the NNRD is the only database to be derived from electronic patient records rather than being a specific, separate collection (Modi et al., 2016).

2.1.2 Contents of the NNRD

The data contained in the NNRD are a defined set of items known as the Neonatal Dataset (NDS), which is an approved NHS Information Standard (SCCI1595). This

means that any neonatal electronic patient record must collect these items as defined, ensuring consistency and future-proofing against change in systems or suppliers. The NDS incorporates core clinical data as it is an electronic patient record, as well as data for commissioning purposes and to determine funding. Established neonatal collections which have gone through considerable clinical review such as the NNAP dataset and the Neonatal Critical Care Minimum Data Set are incorporated. Items in the NDS are therefore generally well-defined and clinically motivated.

While an infant is in neonatal care, data are collected on a static (once per baby), episodic, daily or ad hoc basis. Note that an *episode* refers to a distinct admission to an NNU; if an infant is transferred to another NNU this constitutes a second episode and if the infant subsequently returns to the first NNU this is a third episode. Ad hoc data include occasional events such as screening for retinopathy of prematurity and abdominal x-ray results. There is also a facility to record follow-up data when the infant is two years of age. Variables recorded include infant characteristics (e.g. gestational age and birth weight), parental characteristics (e.g. ethnicity and maternal age), care given (e.g. surgical procedures and respiratory support) and clinical outcomes (e.g. blood stream infections and necrotising enterocolitis). The NNRD currently contains data on over 500,000 infants and 6 million days of care.

2.1.3 NNRD quality assurance

Data input is carried out by variety of NNU staff, mostly nurses and trainee clinicians, although some NNUs employ data clerks. Data are often checked internally by senior clinical staff, and some validation occurs on a wider scale which will be discussed in this section. Utilising data for purposes outside day-to-day clinical care can help motivate improvements data quality. For example activity data feeding into the daily level of care an infant receives is used to determine funding from the NHS and for commissioning services.

Regular data validation occurs through feedback to NNUs of data used in two annual reports: the National Neonatal Audit Program (NNAP) (Royal College of Paediatrics and Child Health, 2014) and the Neonatal Data Analysis Unit (NDAU) annual report for neonatal network mortality (Neonatal Data Analysis Unit, 2015). The NNAP is a national clinical audit established in 2006 to assess consistency of practice and adherence to available guidelines concerning, for example, transfers out of network, screening for retinopathy of prematurity and temperature on admission. For NNAP, quarterly reports to NNUs indicate data completeness, and a final summary is provided for each audit question giving NNUs the opportunity to identify errors or missing data. A similar exercise is conducted annually for the NDAU network mortality report, which includes

data for individual infants, highlights potential errors and gives summary information for each NNU (total number of distinct infants, total number of episodes, total number of deaths occurring in the NNU). In both cases the NNU make corrections themselves on the live data system so there is no direct feedback on the level of errors. The validation process is not obligatory, so does not guarantee that errors are eliminated, and NNUs may be under-resourced in terms of the staff time required to perform the checks.

As part of the MfN programme (Modi et al., 2016) a one-off validation was conducted to compare NNRD data with that recorded independently on the case report form for the Probiotic in Preterm Infants (PiPs) randomised clinical trial. A variety of outcomes, process measures, and infant and maternal characteristics were compared for 1,315 infants born less than 31 weeks gestation who were recruited to the trial between July 2010 and July 2013. Rates of major discordancy were generally low, particularly for data recorded once only at baseline, but there were some exceptions. Individual variables will be discussed in chapter 3 with relevance to the risk factor data used in the main analysis.

2.2 Selection of data for the thesis

In chapter 1 I stated that this thesis will focus on mortality of infants born very preterm (less than 32 weeks gestation). In this section I will further describe the selection of the cohort and definition of the outcome.

2.2.1 Admissions to neonatal care in England

I am only including infants who are admitted to neonatal care. Infants born below 32 weeks gestation who survive birth are always admitted to neonatal care because of the extent of prematurity. These neonatal admissions are a subset of all births below 32 weeks gestation, with the remainder being stillbirths and those who are born alive but die on the labour ward before they can be admitted. Infants who are not admitted to neonatal care are not part of the relevant population for the aims of the thesis. Nevertheless it is important to point out that results cannot be directly compared with estimates of preterm mortality for all births or all live births.

The NNRD in general only contains data on liveborn infants admitted to neonatal units. There are a few exceptions to this as some NNUs contributing to the Vermont Oxford Network Database (an international database of very low birth weight infants born at contributing hospitals) also enter data for infants who died on the labour ward. The scale of the problem is small: eight cases were identified through validation (see 2.1.3) and subsequently excluded. This is less likely to be an issue going forward as from 2014 a new field indicating labour ward deaths was added to the data entry system, so that any NNUs entering data for labour ward deaths could mark them as such, allowing exclusion as necessary.

I will only examine data for infants admitted to English NNUs as data from Wales and Scotland were not available before 2012. Data were contributed from 71% of neonatal units in England from the beginning of 2008, 80% in 2009, 86% in 2010, 97% in 2011, 99% in 2012 and 100% from 2013 onwards. To preserve completeness and restrict analysis to before the further revision of neonatal networks (see chapter 1.2.1) I am only including data for infants born 2011-2013.

2.2.2 Birth weight cohorts

Birth weight has traditionally been a widely used measure with which to identify vulnerable infants. While it is certainly an important predictor of mortality and morbidity in infancy , childhood and later lifer (Barker, 1995; McCormick, 1985), other reasons for

its frequent use may be its ready availability in most settings (including in routine and freely available birth registry data) and the accuracy of measurement (Wilcox, 2001). Standard categories for birth weight are extremely low birth weight (ELBW, less than 1000g), very low birth weight (VLBW, 1000 to 1500g), low birth weight (LBW, 1500 to 2500g), with infants born below 1500g a typical cohort for analysis. At the other end of the scale, infants are considered macrosomic if they have a birth weight greater than 4000 or 4500g, which can also bring risks to the mother and infant (Oral et al., 2001). The remaining range (2500 to 4000 or 4500g) is considered normal range. Infants delivered preterm will generally be smaller so selecting a cohort based on very low birth weight is problematic because the cohort will be heterogeneous, including infants of low gestation along with term infants who are growth-restricted (Arnold et al., 1991). I have therefore defined the cohort based on gestational age. For reference, from Office of National Statistics (ONS) data on live births in below 32 weeks gestation in 2013, 36% were ELBW, 37% were VLBW, 24% LBW and the remainder above 2500g (Office of National Statistics, 2014).

2.2.3 Gestational age criteria

In this thesis, I am focussing on infants born less than 32 weeks gestation for several reasons. Any infant born at this gestation surviving beyond the labour ward would be admitted to a neonatal unit and therefore included in the NNRD. Infants born at later gestations might not be admitted, giving an incomplete denominator. The risk factors for mortality in this population are more established, facilitating risk-adjustment. Finally, mortality rates for this population are more suitable for comparison across neonatal units and networks because amongst infants born at later gestations the rates are so low that precision of the estimates would be very poor.

Hopsitals often have differing policies regarding initiation of active treatment for infants born at very early gestations, leading to variations in mortality (Rysavy et al., 2015; Serenius et al., 2014). In the UK there is no official limit of viability below which resuscitation will not be attempted, but the Nuffield Council for Bioethics (2006, pp. 154-157) proposed the following guidelines:

- before 22 weeks: resuscitation should only take place within a clinical research study.
- 22^{+0} to 22^{+6} weeks: resuscitation should not be standard practice and should only be done following request by the parents and in agreement with clinicians.
- 23^{+0} to 23^{+6}: the decision to resuscitate/give intensive care should be directed by the wishes of the parents, but clinicians should not be obliged to proceed should

they believe the treatment to be futile.

- 24^{+0} to 24^{+6}: standard practice should be full intensive care as required unless the parents and clinicians agree otherwise.
- 25^{+0} and above: standard practice should be full intensive care as required unless the infant is affected by a severe abnormality incompatible with survival.

The British Association of Perinatal Medicine (BAPM) published a framework (Wilkinson et al., 2009) which echoed the Nuffield report, with the additional advice that the response of the heart rate should inform the clinical view.

In this thesis I will only include infants who are admitted to neonatal units, which means that mortality rates could be affected by the decision to resuscitate in the delivery room. There is likely to be variation in resuscitation practices across NNUs due to both clincial and parental views. The EPICure studies (Costeloe et al., 2000, 2012) which examined short and long term outcomes for extremely preterm infants included babies born 22 weeks gestation (i.e. 22^{+0} to 22^{+6}). Some of these infants do survive and they are of particular interest to the neonatal community. Based on this, and the above guidelines, I will exclude infants born before 22^{+0} weeks gestation.

2.2.4 Congenital malformations

Congential malformations are conditions present at birth which are caused by abnormal development of the fetus, and are the second most common cause of infant deaths in England and Wales (Kurinczuk et al., 2010). Infants with congenital malformations are often excluded from studies estimating or comparing neonatal mortality rates (Medlock et al., 2011) either because death is certain, or because the risk of death is substantially higher but the conditions are too rare in the cohort to model this additional risk accurately. Yet there is no agreement in the clinical community about which conditions should be considered lethal (Kurinczuk et al., 2010), and many malformations described as lethal are not always so (Wilkinson et al., 2012). Congenital malformations can be captured in the NNRD using International Classification of Diseases (ICD) codes, a widely-used system maintained by the World Health Organisation. For admitted infants born below 32 weeks gestation in 2008-2014, only 0.1% had a recorded malformation appearing on a list based on a review of the literature published by Wilkinson et al. (2012). Congenital anomaly registers report a rate of around 3% (Kurinczuk et al., 2010), indicating possible underestimation in the NNRD, but the denominator for this figure is all livebirths and stillbirths, and therefore includes terminations and delivery room deaths occurring as a result of the malformation. As the accuracy of the data on

malformations in the NNRD has not been verified I have not attempted to exclude any infants with congenital malformations.

2.2.5 Multiple births

As discussed in section 1.1, multiple birth is one of the risk factors for preterm birth so we would expect a higher frequency of multiple births in this cohort compared with the general population. Of the infants in the NNRD born 2011-2013 below 32 weeks gestation, 73% were singletons, 24% were twins, 2.4% were triplets and 0.2% were quadruplets or higher. Infants from the same pregnancy may have similar clinical outcomes due to common factors such as genetic traits, lifestyle of the parents, and circumstances around the pregnancy and birth (Ananth et al., 2005; Carlin et al., 2005; Yelland et al., 2015). The consequences of correlated outcomes for statistical analysis in general was discussed briefly in chapter 1 and will be addressed further in chapter 4 in the context of provider profiling. Examining the extent to which outcomes are correlated requires matching infants from the same pregnancy (multiple birth sets). Multiple birth sets can be linked in the NNRD using a maternal identification number, which is usually an encrypted version of the mother's NHS number. Mother's NHS number was missing for around 11% of infants; in these cases a maternal ID was assigned during data processing conducted by NDAU. Infants are assigned to the same maternal ID if the following variables are identical: month and year of birth[1], gestational week at birth, place of birth (in most cases, hospital identifier), maternal year of birth and area of residence (Statnikov and Gray, 2015).

A substantial proportion of multiple birth sets are missing at least one sibling in the NNRD: 26% of twin sets, 39% of triplet sets and 48% of quadruplet or higher sets (for infants born 2011-2013 below 32 weeks gestation). As described in setion 2.2.1 very preterm infants who survive beyond the labour ward are always admitted to neonatal care. Siblings could be missing because they die (in-utero or postnatally) before they can be admitted to a neonatal unit. The definition of multiple birth in the NNRD is for the total number of fetuses delivered, whether live or stillborn. There is no specific field in the NNRD for recording the death of a sibling from the same pregnancy.

It is possible that the missing siblings are in the NNRD, but there is insufficient information to link them. This seems a likely contributing factor given that the proportion of multiple births with missing siblings was 12% in those where mother's NHS number was available, compared with 43% in those where it was missing. The sibling could have been admitted to neonatal care but the record is missing from the NNRD. Infants in multiple birth sets are sometimes separated due to insufficient cot capacity

[1]exact date is removed as part of pseudonymisation

(Gale et al., 2012b). The infant with no record may have been transferred to another neonatal unit not contributing to the NNRD, but postnatal transfers are usually still recorded as a neonatal admission in the unit of birth. Furthermore at least 97% of neonatal units were contributing data in 2010 onwards so the likelihood of admission to a non-contributing unit is low. Another possibility is that the infant with a present record was erroneously recorded as being from a multiple pregnancy, but this cannot be determined using NNRD data alone.

To investigate whether mortality within multiple birth sets is correlated, I examined outcomes for the 3625 twin pairs that could be matched (7250 infants in total). The overall mortality rate was 8.8%, but if one sibling died, the mortality rate amongst the second siblings was 38.5%. Some of this correlation will be accounted for by the fact that the twins are of the same gestation, and other factors contributing to the similarity could be included as risk factors (e.g. maternal smoking), but there is likely to be considerable residual correlation from genetic effects and other unmeasured similarities (Carlin et al., 2005).

The aim of this thesis is to develop models to estimate NNU mortality rates accounting for the structural features of the data imposed by centralised care. Multiple births impose a separate structure on the data, and are important given the high proportion of multiple births in the population. However this is a separate issue to the one currently being addressed, one which brings with it a separate set of problems, not least because of the high proportion of missing siblings in the data set. I will therefore only consider singleton infants for analysis in this thesis, including the descriptive analysis presented later in this chapter (section 2.3). The limitations of this will be discussed in section 2.4 and I will consider the potential for extension to include multiple births in chapter 9.

2.2.6 Defining mortality

Mortality was selected as the outcome because it is unambiguous, clearly defined, well-recorded, and of the utmost importance to clinicians and parents. Determining which deaths to include can complicate comparative analyses of mortality of preterm infants. For international comparisons that can be particularly difficult because of variations in stillbirth definition, registration of births, limits of viability and resuscitation policies (Draper et al., 2009; Joseph et al., 2012). In this thesis I am only considering comparison of mortality rates within England but there are still a number of issues which require consideration.

Timing of deaths

Neonatal mortality is usually defined as death within 28 days, and sometimes subdivided into *early neonatal mortality* (within 7 days) and *late neonatal mortality* (between 7 and 28 days). Globally, over a third of deaths in children under 5 occur within the first 28 days of life (Lawn et al., 2005). These time points do achieve standardisation but they are somewhat arbitrary. As neonatal care has progressed, an increased number survive the neonatal period of 28 days but some of these infants die later during stay (Sidebotham et al., 2014). Applying the 28 day definition means those infants would be considered survivors. Furthermore, for the purposes of provider profiling all deaths which occur in neonatal care are of interest. I will consider death before discharge from neonatal care as the outcome in this thesis to ensure all deaths occurring on an NNU are included, but death before 28 days will be included as part of the description of the cohort in section 2.3.4 for comparison with other data sources.

Hospital mortality measures sometimes include deaths up to 30 days post-discharge (e.g. the Summary Hospital Mortality Index produced for each English NHS Trust (Health and Social Care Information Centre Clinical Indicators Team, 2016)), though this is mainly relevant for mortality following elective procedures. While some work has been done to link infants in the NNRD with death registry data (via the Office of National Statistics and Hospital Episode Statistics linked mortality dataset) this was only possible for a limited number of infants so was not considered further.

2.2.7 Missing mortality data

Mortality was determined by the discharge record for the last episode of care. If the last discharge was a transfer for further neonatal care with no subsequent data available, death before discharge was considered missing; this occurred for 2% of infants (341/16,710 based on singleton infants born 2011-2013 at or below 32 weeks gestation). A similar level of missing outcomes persists even beyond 2013 when all NNUs contributed data. Examination of infants with missing outcomes showed that most were transferred to specialist surgical providers that are not standard neonatal units and do not contribute data to the NNRD, such as Great Ormond Street Hospital. While these centres are not NNUs, infants are transferred there and may be transferred back to an NNU after the procedure, so the infants are still considered to be in neonatal care for the duration. Infants not transferred back may be discharged home or to a paediatric ward, or they may die at the specialist centre. Table 2.1 compares infant characteristics for those with and without death data. Infants with missing death data tended to be more vulnerable: of lower gestation, lower birth weight and more likely to be male (risk factors for mortality will be described in more detail in chapter 3). They were also more

likely to have missing data in other fields.

In order to assess what impact the missing data might have on the results we need to consider why it is missing. Following Rubin (1976), missing data can be classified in three ways: *missing completely at random* (MCAR), *missing at random* (MAR) or *missing not at random* (MNAR). If data are MCAR, this means that the probability of the data being missing is completely unrelated to the data itself, either observed or unobserved (e.g. if some records are randomly deleted due to a computer error). In such cases analysis of the complete case data only will give unbiased results, though with larger standard errors than if the full dataset were available due to the sample size reduction. The MAR criteria are less stringent than MCAR: here the probability of missingness is conditionally independent of the unobserved data, given the observed data. An example of this would be if men were less likely to fully complete a health questionnaire than women, but this is unrelated to the answers beyond any gender differences in the observed data. We can obtained unbiased inferences provided we fill in (*impute*) or re-weight the variables with missing data conditional on what we observe (Little, 1987). If the missingness is dependent on the data which are missing (e.g. if those with poor health are less likely to fill out the questionnaire in full but we have no data relating to health status) then the data are MNAR. Here we have to model the probability of missingness as well as the analysis model of interest. As the missingness mechanism is usually unknown, sensitivity analysis is crucial to check whether results are robust to a variety of assumptions.

For infants with missing outcomes in the NNRD we can rule out MCAR because most are transferred to particular providers for specialist treatment, and so differ systematically from other infants. Missing at random is an untenable assumption as the missingness is related to the outcome. Infants that die in the specialist centre will certainly have a missing outcome in the NNRD, while those that survive may have a missing outcome if they are discharged home or to a paediatric ward. We cannot impute the missing outcomes based on the subpopulation of infants who were transferred for specialist care for whom we do know the outcome as these infants are transferred back to an NNU and therefore exclude deaths at the specialist centre. It therefore seems most appopriate to treat the data as MNAR and further understanding of the missing death data is required in order to construct sensible models for the missingness mechanism.

Given these complexities, for the purposes of exploring the main statistical issues at hand I have limited the analysis in this thesis to complete outcome data, and acknowledge that this may bias the results given the reasons for missingness and the differences in characteristcs observed in table 2.1. Inclusion of infants with missing outcomes will be discussed in chapter 9.

Table 2.1: Comparison of infants with and without missing death data in the NNRD, singleton infants born less than 32 weeks gestation 2011-2013

		Non-missing		Missing		Total	
		n	%	n	%	n	%
Gestational age (weeks$^{+\text{days}}$)	$22^{+0} - 23^{+6}$	398	2.4	13	3.8	411	2.5
	$24^{+0} - 25^{+6}$	1873	11.4	76	22.3	1949	11.7
	$26^{+0} - 27^{+6}$	2896	17.7	82	24.0	2978	17.8
	$28^{+0} - 29^{+6}$	4443	27.1	90	26.4	4533	27.1
	$30^{+0} - 31^{+6}$	6759	41.3	80	23.5	6839	40.9
Birth weight (g)	< 500	131	0.8	9	2.6	140	0.8
	$500 - 999$	5378	32.9	182	53.4	5560	33.3
	$1000 - 1499$	6839	41.8	95	27.9	6934	41.5
	$1500 - 1999$	3725	22.8	50	14.7	3775	22.6
	$\geqslant 2000$	296	1.8	5	1.5	301	1.8
Sex	Female	7342	44.9	135	39.6	7477	44.7
	Male	9027	55.1	206	60.4	9233	55.3
Mode of delivery	Vaginal delivery	6793	41.5	136	39.9	6929	41.5
	Caesarean section	8377	51.2	152	44.6	8529	51.0
	Missing	1199	7.3	53	15.5	1252	7.5
Maternal age (years)	<20	1327	8.1	29	8.5	1356	8.1
	20 to 24	3166	19.3	71	20.8	3237	19.4
	25 to 29	4292	26.2	84	24.6	4376	26.2
	30 to 34	4336	26.5	92	27.0	4428	26.5
	35 to 40	2451	15.0	49	14.4	2500	15.0
	>40	726	4.4	9	2.6	735	4.4
	Missing	71	0.4	7	2.1	78	0.5
Maternal	White	10909	66.6	215	63.0	11124	66.6
	Mixed	271	1.7	9	2.6	280	1.7
	Asian	2119	12.9	35	10.3	2154	12.9
	Black	1702	10.4	29	8.5	1731	10.4
	Chinese	55	0.3	3	0.9	58	0.3
	Other	248	1.5	5	1.5	253	1.5
	Missing	1065	6.5	45	13.2	1110	6.6
Smoking*	No	10605	64.8	199	58.4	10804	64.7
	Yes	3145	19.2	64	18.8	3209	19.2
	Missing	2619	16.0	78	22.9	2697	16.1
IMD† quintile	1 (most deprived)	5543	33.9	115	33.7	5658	33.9
	2	3695	22.6	79	23.2	3774	22.6
	3	2740	16.7	54	15.8	2794	16.7
	4	2081	12.7	39	11.4	2120	12.7
	5 (least deprived)	1708	10.4	35	10.3	1743	10.4
	Missing	602	3.7	19	5.6	621	3.7

* Any smoking in pregnancy

† Quintile of the ranked Index of Multiple Deprivation score for the mother's Lower Super Output Area

2.3 Descriptive analysis

2.3.1 Infant characteristics

Table 2.2 shows the population characteristics for infants born 2011-2013 below 32 weeks gestation and admitted to an NNU in England contributing to the NNRD. Infants with a birth weight greater than 4 standard deviations from the gestation and sex specific mean (based on UK-WHO preterm standards (Wright et al., 2010)) were excluded (n=23) as gestation may have been incorrectly recorded. As discussed in section 2.2.7, those with missing death data were excluded. All subsequent analyses in this chapter are based on the cohort described in table 2.2.

The number of infants increases with gestational age, with 398 (2.4% of the cohort) infants born at 22^{+0} to 23^{+6} $\text{weeks}^{\text{days}}$ gestation over the three year period increasing to 6759 (41.3%) at 30^{+0} to 31^{+6}. A quarter of infants had a birth weight of over 1500g, highlighting differences between gestational age and birth weight cohorts. There were more males than females, slightly higher than for live births across all gestations (55.1% vs 51.2% (Department of Health, 2014)). Half the infants were born by Caesarean section compared with 13.2% of all deliveries. The proportion of infants born to white mothers was lower than for live births overall (Office of National Statistics, 2015) (66.6% vs 73.0%). The proportion of infants born to black mothers was higher (10.4% vs 4.5%), as was the proportion born to Asian mothers, but to a lesser extent (12.9% vs 9.2%). Nineteen percent of mothers were recorded as having smoked during pregnancy, which is the same as the overall smoking prevalence amongst adults (Health and Social Care Information Centre, 2016a), though almost as many were missing so this is likely to be an underestimate. There is a disproportionate representation of deprived areas with a third of infants coming from the fifth most deprived areas[2].

Table 2.2 also shows data by transfer status, and mortality data; these will be discussed in sections 2.3.3 and 2.3.4 respectively.

[2]Quintiles are determined by ranking Lower Super Output Areas (LSOAs) according to Index of Multiple Deprivation 2010 (IMD) score. LSOAs are geographic areas approximately equal in population size and are determined using mother's postcode. The IMD is a composite score assessing deprivation based on income, employment, health, education, housing, crime and living environment calculated by the UK Department of Communities and Local Government (Department for Communities and Local Government, 2011).

Table 2.2: Description of cohort for the thesis (singleton infants born less than 32 weeks gestation 2011-2013, admitted to a NNU in England contributing to the NNRD, with known survival status at discharge), by transfer status

		Non-transferred		Transferred		Total	
		n	%	n	%	n	%
Gestational age (weeks$^{+\text{days}}$)	$22^{+0} - 23^{+6}$	223	2.2	175	2.9	398	2.4
	$24^{+0} - 25^{+6}$	701	6.8	1172	19.2	1873	11.4
	$26^{+0} - 27^{+6}$	1394	13.6	1502	24.6	2896	17.7
	$28^{+0} - 29^{+6}$	2813	27.4	1630	26.7	4443	27.1
	$30^{+0} - 31^{+6}$	5129	50.0	1630	26.7	6759	41.3
Birth weight (g)	< 500	68	0.7	63	1.0	131	0.8
	$500 - 999$	2464	24.0	2914	47.7	5378	32.9
	$1000 - 1499$	4626	45.1	2213	36.2	6839	41.8
	$1500 - 1999$	2879	28.1	846	13.8	3725	22.8
	$\geqslant 2000$	223	2.2	73	1.2	296	1.8
Sex	Female	4691	45.7	2651	43.4	7342	44.9
	Male	5569	54.3	3458	56.6	9027	55.1
Mode of delivery	Vaginal delivery	4112	40.1	2681	43.9	6793	41.5
	Caesarean section	5321	51.9	3056	50.0	8377	51.2
	Missing	827	8.1	372	6.1	1199	7.3
Maternal age (years)	<20	851	8.3	476	7.8	1327	8.1
	20 to 24	1967	19.2	1199	19.6	3166	19.3
	25 to 29	2746	26.8	1546	25.3	4292	26.2
	30 to 34	2731	26.6	1605	26.3	4336	26.5
	35 to 40	1491	14.5	960	15.7	2451	15.0
	>40	426	4.2	300	4.9	726	4.4
	Missing	48	0.5	23	0.4	71	0.4
Maternal ethnicity	White	6801	66.3	4108	67.2	10909	66.6
	Mixed	168	1.6	103	1.7	271	1.7
	Asian	1377	13.4	742	12.1	2119	12.9
	Black	1025	10.0	677	11.1	1702	10.4
	Chinese	32	0.3	23	0.4	55	0.3
	Other	155	1.5	93	1.5	248	1.5
	Missing	702	6.8	363	5.9	1065	6.5
Smoking*	No	6438	62.7	4167	68.2	10605	64.8
	Yes	2065	20.1	1080	17.7	3145	19.2
	Missing	1757	17.1	862	14.1	2619	16.0
IMD† quintile	1 (most deprived)	3677	35.8	1866	30.5	5543	33.9
	2	2268	22.1	1427	23.4	3695	22.6
	3	1622	15.8	1118	18.3	2740	16.7
	4	1238	12.1	843	13.8	2081	12.7
	5	1060	10.3	648	10.6	1708	10.4
	Missing	395	3.8	207	3.4	602	3.7
Death before discharge	Survived	9141	89.1	5584	91.4	14725	90.0
	Died	1119	10.9	525	8.6	1644	10.0
Time of death	Died within 24 hours	393	35.3	25	4.8	418	25.6
	Died 1-7 days	300	27.0	123	23.6	423	25.9
	Died 8 to 28 days	264	23.7	167	32.0	431	26.4
	Died after 28 days	155	13.9	207	39.7	362	22.2

* Any smoking in pregnancy

† Quintile of the ranked Index of Multiple Deprivation score for the mother's Lower Super Output Area

2.3.2 Structure of neonatal care

In section 1.2 I described the different levels of neonatal unit (special care baby unit (SCBU), local neonatal unit (LNU), and neonatal intensive care unit (NICU) in order of increasing levels of care provided. This section describes the relative burden of care for these NNUs in terms of patient volume and type of care.

For 2011-2013 there were data from 169 neonatal units (45 SCBUs, 78 LNUs and 46 NICUs) in 23 neonatal networks (see appendix A for a full list). Table 2.3 summarises the number of admissions (episodes of care) and the number of distinct infants (i.e. excluding multiple admissions of the same infant to the same NNU) by level of neonatal unit for the cohort described in section 2.3.1. Figure 2.1 illustrates the distribution of gestational age of admissions (not infants, so each admission counts and transferred infants are therefore included multiple times). The higher the level of NNU, the more infants are admitted and the lower gestational age. There is considerable variation in care volumes even within NNU level.

	SCBU	LNU	NICU
No. NNUs	45	78	46
No. admissions	61 (41, 83)	117 (86,149)	266 (226,325)
No. infants	47 (31, 62)	97 (74,134)	242 (206,290)
Length of stay (days)	14 (11,20)	30 (28,33)	34 (31,36)

Table 2.3: Summary of admissions by level of NNU for singleton infants born less than 32 weeks gestation 2011-2013; data shown as median (interquartile range) across NNUs

	SCBU	LNU	NICU
Number of IC days	22 (13, 42)	402 (240, 645)	2558 (2016,3619)
Percentage of IC days	1.9 (1.3,3.1)	10.6 (8.2,15.1)	28.0 (25.2,32.7)
Number of HDC days	87 (29, 286)	892 (565,1285)	2896 (2310,3874)
Percentage of HDC days	7.1 (3.4,14.8)	24.1 (18.9,27.8)	30.6 (27.0,33.7)
Number of SC days	1152 (747,1506)	2364 (1838,3258)	3788 (2857,4968)
Percentage of SC days	91.2 (84.7,95.3)	64.6 (57.8,72.3)	40.8 (34.8,45.5)

Table 2.4: Summary of type of care by level of NNU for singleton infants born less than 32 weeks gestation 2011-2013; data shown as median (interquartile) across NNUs

In section 1.2 I described how each day that an infant is in neonatal care is classified as special care (SC), high dependency care (HDC) and intensive care (IC) in order of intensity. These categories form an integral part of how NNUs are funded so data relating to the definitions are likely to be well-collected (Department of Health, 2009; Gale et al., 2016b). Table 2.4 summarises the number and proportion of care days in each category across NNUs, by level of NNU for the cohort described in section 2.3.1. Higher level NNUs perform higher levels of care, and greater proportions of it with

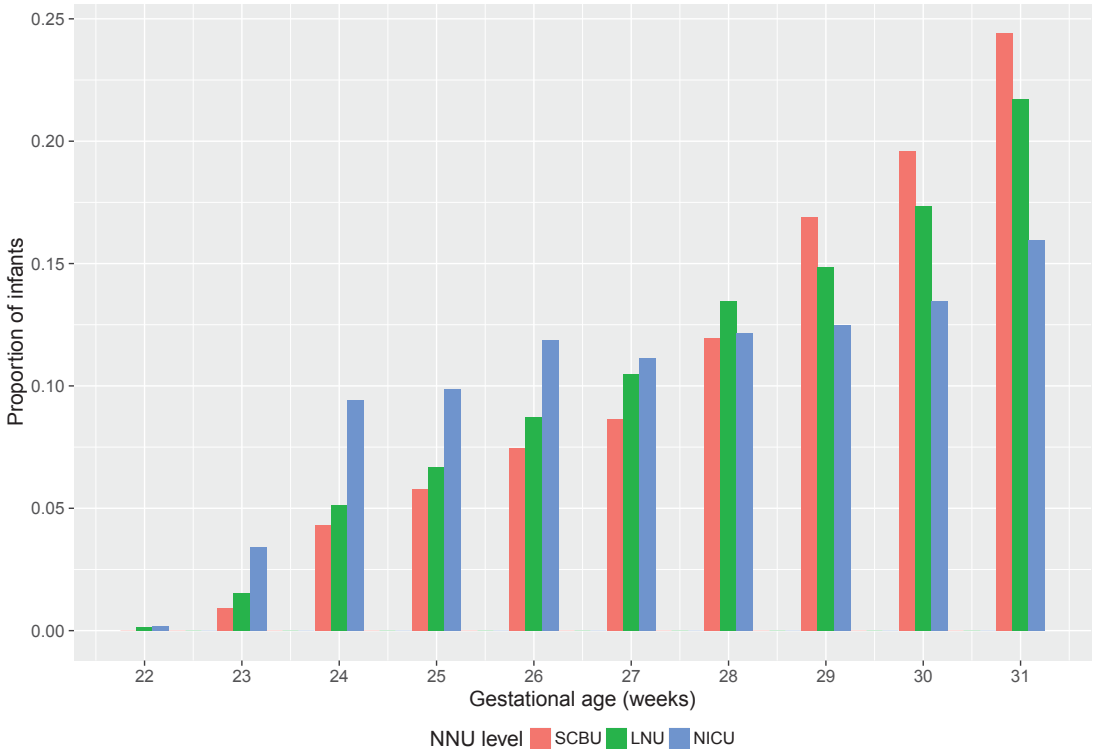


Figure 2.1: Distribution of gestational age at birth by NNU level of admission, singleton infants born below 32 weeks gestation 2011-2013

some variation within NNU level. Although special care constitutes the majority of care carried out at SCBUs (median 91% of care days), NICU conduct more special care in terms of volume (median 3788 vs 1152 days). To give an indication of what category of care happens when and where, figure 2.2 shows the number of care days by category and postnatal age separately for each level of NNU. SCBUs have a peak in volume on day one, on which over half the care days are intensive care, followed by a drop as infants are transferred out to higher level NNUs. The volume of care then increases again, and consists almost wholly of special care, as infants are transferred back to their nearest NNU. In contrast, NICUs have a volume peak a few days after birth as infants are transferred in and the majority of care for the first ten days is intensive care. The pattern for LNUs is in between that for SCBUs and NICUs, with a drop after day one but a substantial amount of intensive care carried out in the first few days. By day 28 nearly all intensive care is conducted in NICUs.

2.3.3 Transfers

Neonatal transfers were introduced in section 1.2.2 and reasons for transfer were described. Transfers can occur either before or after birth, but whether in-utero transfer

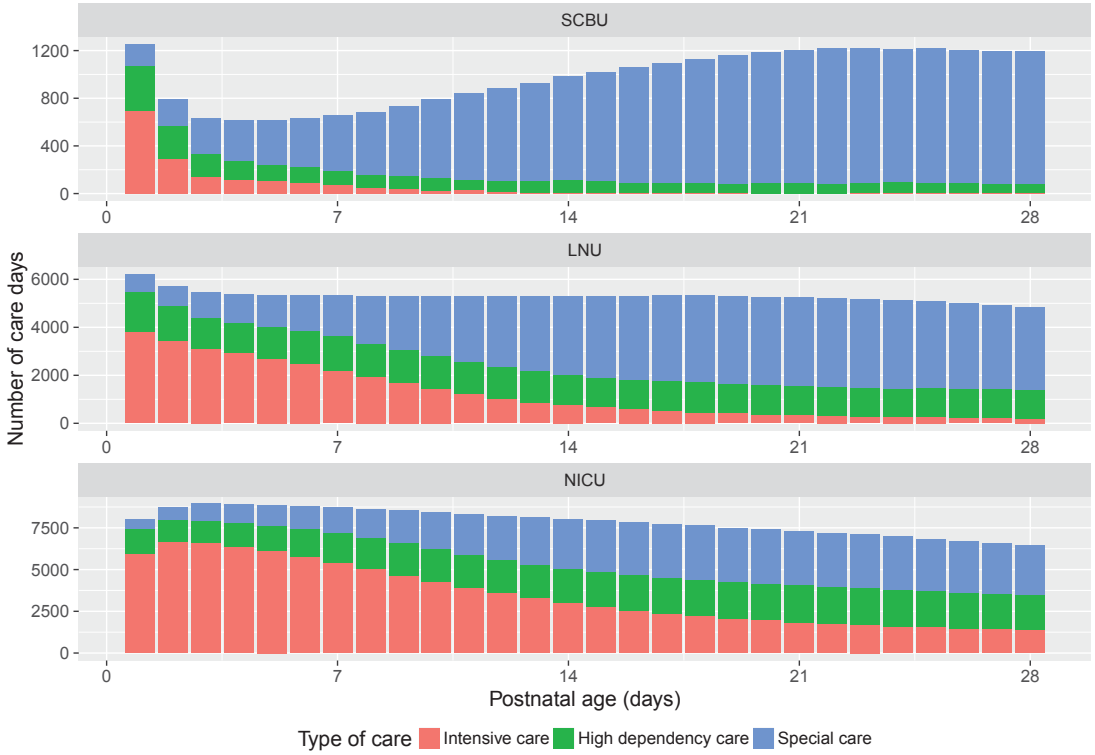


Figure 2.2: Number of care days at each category, by postnatal age and NNU level, singleton infants born below 32 weeks gestation 2011-2013; scales are not comparable to show patterns at SCBUs more clearly

is carried out or attempted is not recorded in the NNRD. In this thesis I am primarily interested in transfers from the point of view of care provided by multiple NNUs, and so only postnatal transfers are considered.

For each episode the reasons for admission and discharge are recorded. This information could be used to classify transfers, though the admission and discharge fields do not have the same categories (note they also do not correspond directly with the types of transfer described in section 1.2.2). Table 2.5 compares the reason for admission for transferred infants with the reason for discharge from the previous episode. If an infant is transferred we would expect these reasons to correspond, but the extent to which this occurred was limited. For example 60% of discharges for surgical reasons were recorded as admissions for surgical reasons, and around 200 episodes were recorded as ending with the infant being discharged home, even though a subsequent admission was recorded. I therefore decided not to use these fields to classify transfers as they are unreliable, but instead described transfers based on the levels of the discharging and admitting NNUs. The timing of transfer and whether it is to a higher, lower or same level NNU can tell us something about the type of transfer. For example early transfers to a higher level NNU are likely to be acute transfers while later transfers to a lower level NNU are likely

to be non-acute transfers.

It is difficult to determine whether a transfer occurred for capacity reasons because this depends not only on whether there is physical cot available but also on staffing, both of which cannot be determined using the NNRD[3].

	Reason for previous discharge						
Reason for admission	Home/ foster care	Paediatric ward	Continuing care	Specialist care	Surgical care	Cardiac care	Unknown/ Missing
IC	62	8	1933	1007	196	32	3
HDC	10	1	1236	62	30	2	1
SC	47	8	1712	49	18	2	0
Surgical care	46	0	65	134	515	2	2
Cardiac care	0	0	8	17	38	58	0
Specialist investigation	6	0	7	62	14	1	0
Back transfer for IC	0	1	323	5	7	11	1
Back transfer for HDC	3	1	529	5	3	0	0
Back transfer for SC	2	2	935	10	4	2	0
Social care	0	0	4	0	0	0	0
Transitional care	4	4	42	1	0	0	0
Other	16	0	8	20	13	1	0
Unknown/Missing	5	0	49	19	17	1	0

Table 2.5: Comparison of reasons for transfer with reason for admission for subsequent episode, singleton infants born below 32 weeks gestation 2011-2013

Frequency Of the cohort described in section 2.3.1, 37% (6,109/16,369) were transferred at least once. Half of transferred infants (3,035/6,109) required only one transfer and while the maximum number of transfers was 12, this was very unusual (around 1% had 6 or more transfers).

Comparison of transferred and non-transferred infants Table 2.2 showed that the gestational age and birth weight distributions differed for transferred and non-transferred infants, with the latter more likely to be of higher gestation and birth weight. The proportion of transferred infants born at less than 28 weeks gestation was over double that for non-transferred infants (46.7% vs 22.6%). Similarly, the proportion of transferred infants born at less than 1000g was almost double that for non-transferred infants (48.7% vs 24.7%). The proportion of boys was slightly higher for transferred infants (56.6% compared with 54.3%). While transferred infants were less likely to come from the most deprived quintile, this was not part of a trend reflected in the remaining quintiles. Mortality differences between transferred and non-transferred infants will be discussed in section 2.3.4

[3]Admission month and year are available in the NNRD but exact admission dates are not (all events including admission times are recorded in minutes from birth). Therefore it is not possible to establish the number and identity of the infants in an NNU at a particular time. Staffing data are not recorded.

Gestational age The proportion of infants transferred is highest at the middle of this gestational age range (figure 2.3). Infants born before the peak at 25 weeks are no more likely to be born at hospital with a NICU (which remains at around 70% up to 27 weeks). Mortality rates are much higher below 25 weeks so it is possible that these infants die before a transfer would have otherwise occurred. As gestational age increases the transfer rate decreases, and infants are less likely to require multiple transfers also. This is likely to be because of a reduced need for transfer as these infants will not necessarily require care at a NICU (British Association of Perinatal Medicine, 2010).

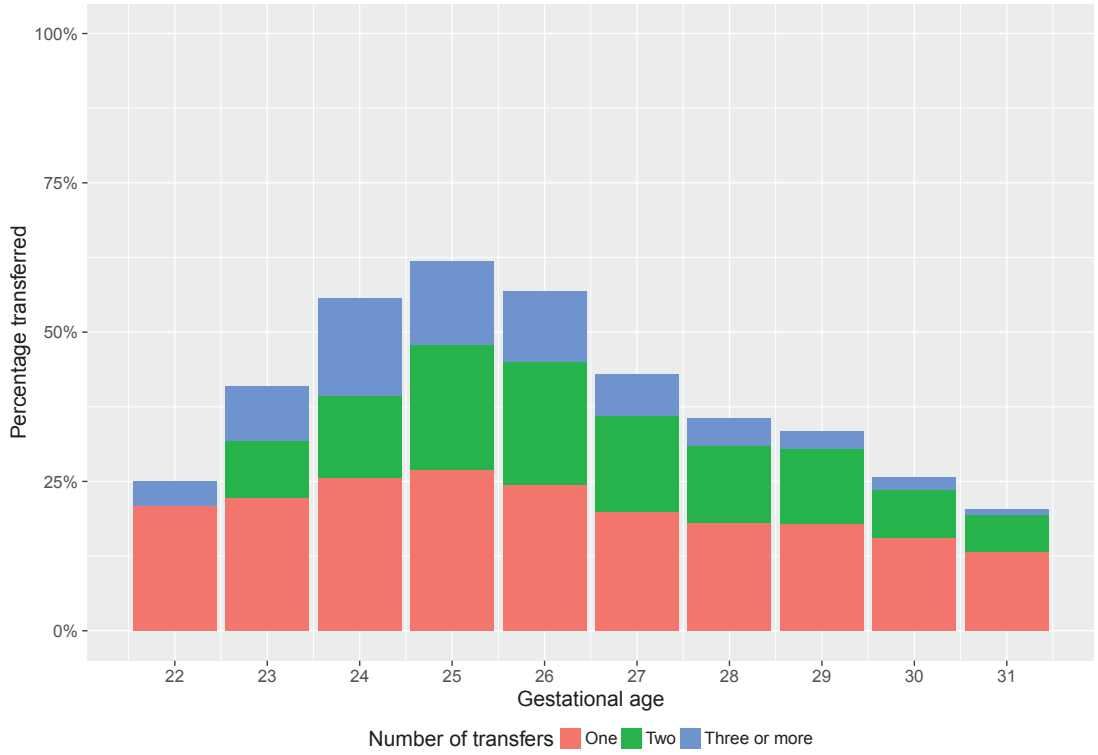


Figure 2.3: Percentage of infants transferred at each gestational age week, grouped by the number of transfers, for singleton infants born below 32 weeks gestation 2011-2013

NNU level The median (interquartile range, IQR) across NNUs for the percentage of infants transferred at least once was 79% (67,91) for SCBUs, 47% (39,54) for LNUs and 53% (44,63) for NICUs. These percentages are based on unique infants admitted to each NNU, so each transferred infant will contribute to every NNU he/she is treated at, but repeat admissions to the same NNU are only counted once. Figure 2.4 shows the level of the destination NNU by the level of the referral NNU. Note that while there are fewer transfers from SCBUs, only 12% of episodes of care for this cohort occur at SCBUs, 30% of which end in transfer.

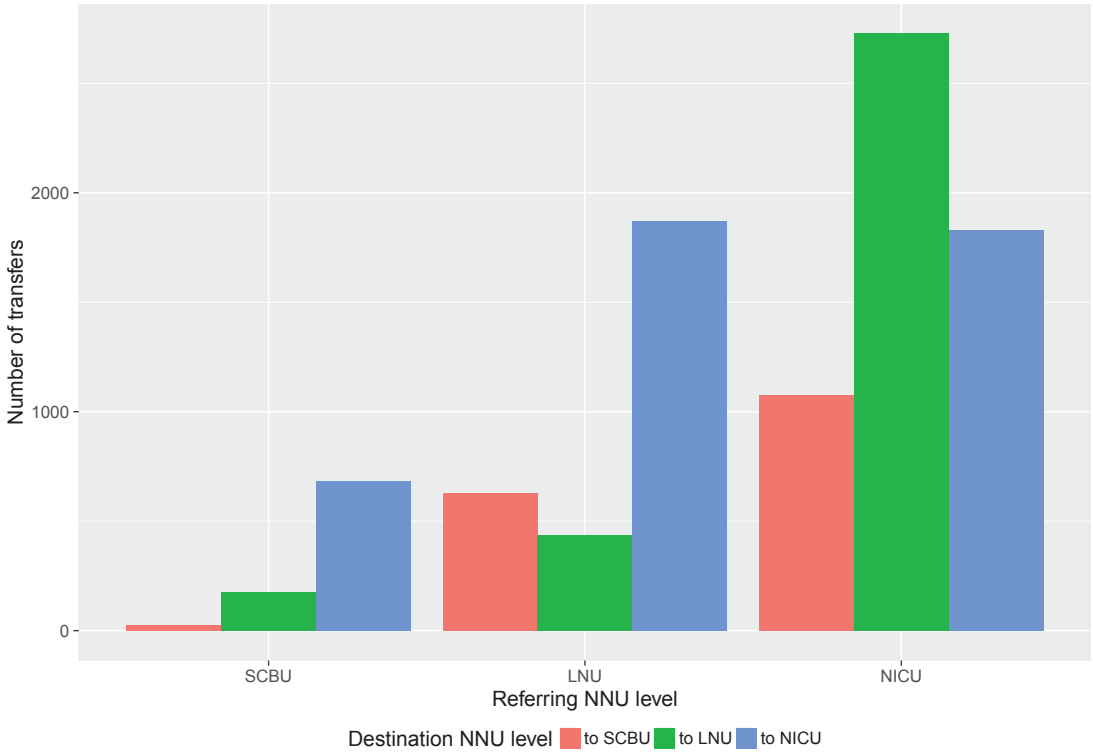


Figure 2.4: Direction of neonatal transfers by level of the referring NNU, singleton infants born below 32 weeks gestation 2011-2013

Timing Figure 2.5 shows the timing of transfers, by whether they are to a lower, same or higher level NNU. A large number of transfers occur within the first 24 hours after birth, with most of these being to a higher level NNU (figure 2.5). These are likely to represent acute transfer of infants born at SCBUs or LNUs but who require specialist treatment at a NICU; figure 2.4 shows that NICUs are the destination of most transfers from SCBUs and LNUs. There are still some transfers to a higher level occurring even at much later postnatal ages. Transfers to a lower level make up an increasing proportion of the total transfers over time. These are likely to be non-acute transfers back to the NNU closest to home, once the infant's needs are less complex. However there are also a few transfers to a lower level on day one. There are a non-negligible number of transfers to another NNU of the same level, particularly after the first week, most of which represent NICU to NICU transfers (figure 2.4).

Cross-network transfers In the cohort 26% (2147/9542) of transfers were to neonatal unit in a different network. The frequency of transfers within and between networks is shown in figure 2.6. The diagonal points represent within-network transfers, with some networks (e.g. London North East and Greater Manchester) transferring more frequently. Networks are plotted in order of region and adjacent regions plotted to-

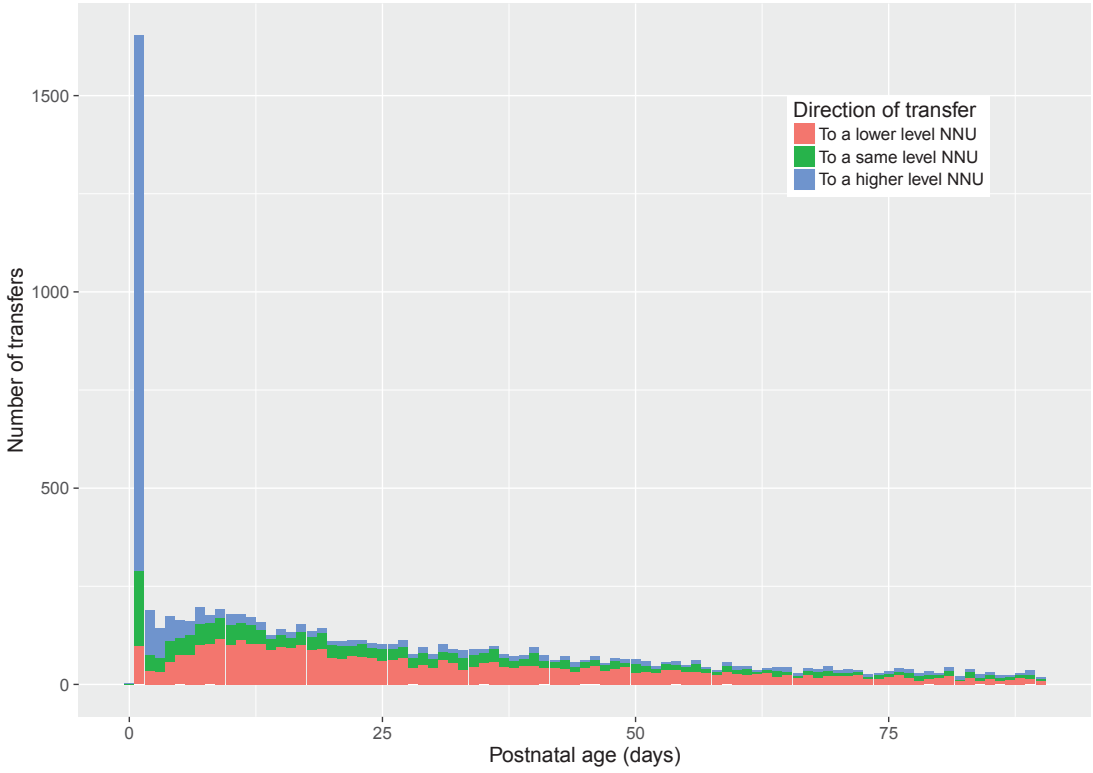


Figure 2.5: Timing of neonatal transfers, grouped by direction of transfer for singleton infants born below 32 weeks gestation 2011-2013

gether where possible. Most points are close to the diagonal line indicating that most cross-network transfers are to a nearby network. Off-diagonal points are numerous but generally quite small. This indicates that cross-network transfers are spread amongst serveral different networks with a few transfers each, rather than mostly to one other network. There are some networks which tend to transfer more to each other, for example London South East and Kent. The plot is fairly symmetrical, indicating that the number of transfers between pairs of NNUs is similar in both directions (an example exception to this that the Western network transfers more infants to Midlands South West than it receives). Of the transfers to a different network, 39% (946/2147) were to another NNU of the same level (mostly NICU to NICU) in comparison with 19% of the within-network transfers.

2.3.4 Mortality

The overall mortality rate was 10%, with around a quarter of deaths occurring within 24 hours of birth, another quarter by 7 days, another quarter by day 28 and another quarter after 28 days (table 2.2). Associations between mortality and infant characteristics will be discussed in chapter 3.

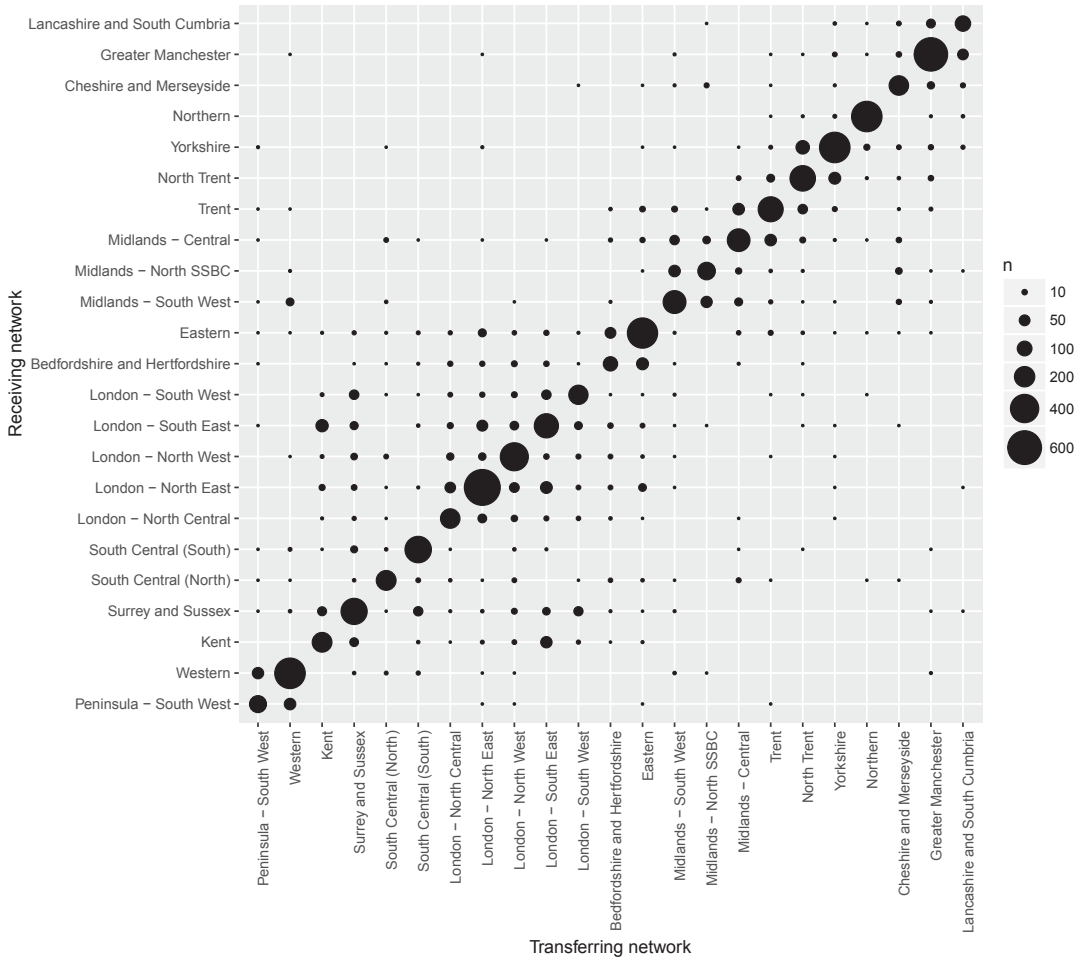


Figure 2.6: Frequency of transfers between neonatal networks; area of the dot is proportional to the number of transfers and networks ordered by region; singleton infants born below 32 weeks gestation 2011-2013

Table 2.2 shows how mortality differs by transfer status. Mortality was lower amongst transferred infants compared with non-transferred infants (8.6% vs 10.9%) and the timing of deaths was very different for these groups. Few deaths occurred within 24 hours of birth for the transferred infants because they have to survive long enough to be transferred (4.8% vs 35.3% in non-transferred infants), and more occurred after 28 days (39.7% vs 13.9%). Table 2.6 illustrates differences in mortality by transfer status and NNU level. Mortality was calculated for each NNU as the mortality rate amongst all infants who ever received care in the neonatal unit. Non-transferred infants are therefore included for only one NNU, but transferred infants (who also appear in the total) will be included at all NNUs caring for them, as indicated by the asterisks in table 2.6. Higher level NNUs had higher crude mortality rates, with the biggest differences amongst infants who were not transferred (median mortality at SCBUs 2.6%, LNUs 5.2%, NICUs 15.1%). The aim of examining distribution of deaths was to illustrate where the deaths actually occur, so deaths of transferred infants were attributed

to the final NNU. The vast majority of deaths occur in NICUs, whether for transferred (93.3%) or non-transferred infants (76.1%). In contrast, 39.3% of all infants who survive are discharged from a NICU; this is lower (22.8%) for those who are transferred. These results indicate how different the care pathways are for infants who survive and those that die. Mortality was also expressed overall by NNU level, again allocating infants to the final NNU. Mortality was lower for transferred infants than non-transferred infants for SCBUs and LNUs, and higher for NICUs. This is opposite to differences in median NNU mortality where infants are allocated to all NNUs.

Table 2.6: Mortality across levels of NNU, singleton infants born below 32 weeks gestation 2011-2013

	SCBU	LNU	NICU
All infants			
Mortality across NNUs %, median(IQR)*	4.7 (3.2,6.8)	6.1 (4.4,8.1)	12.2 (9.3,15.5)
Distribution of deaths, n(%)†	33 (2.0)	269 (16.4)	1342 (81.6)
Distribution of survivals to discharge, n(%)†	2195 (14.9)	6738 (45.8)	5792 (39.3)
Overall mortality %†	1.5	3.8	18.8
Non-transferred infants			
Mortality across NNUs %, median(IQR)	2.6 (0.0,11.2)	5.2 (2.8,7.7)	15.1 (10.9,21.1)
Distribution of deaths, n(%)	29 (2.6)	238 (21.3)	852 (76.1)
Distribution of survivals to discharge, n(%)	545 (6.0)	4075 (44.6)	4521 (49.5)
Overall mortality %	5.1	5.5	15.9
Transferred infants			
Mortality across NNUs %, median(IQR)*	4.7 (3.0,7.1)	8.3 (4.7,10.6)	9.8 (8.4,11.6)
Distribution of deaths, n(%)†	4 (0.8)	31 (5.9)	490 (93.3)
Distribution of survivals to discharge, n(%)†	1650 (29.5)	2663 (47.7)	1271 (22.8)
Overall mortality %†	0.2	1.2	27.8

* attributed to all NNUs caring for the infant
† Infants attributed to the final NNU

2.4 Discussion

2.4.1 Summary

In this chapter I described the data source for the thesis (the NNRD), the selection of the cohort, and presented descriptive analysis of the data. The NNRD is constructed from electronic clinical records for all admissions to neonatal care in England, Wales and Scotland and contains a wide range of data relating to demographics, clinical observations, treatments and outcomes. I selected singleton infants born 2011-2013 between 22^{+0} to 31^{+6} weeks gestational age who were admitted to neonatal care in England. The outcome was selected as death (from any cause) before discharge from neonatal care, and any infants with missing death data were excluded. Description of infant characteristics showed that the number of infants increased with gestational age. Males, infants born to black mothers, and infants born to mothers from more deprived areas were over-represented compared with national statistics for births across all gestations.

Higher level NNUs perform more days of care and at a higher category of care overall, and on a per-NNU basis, but there was considerable variation within levels of NNU. The intensity of care decreased with postnatal age. Over a third of infants were transferred at some point during stay, many of which occurred during the first 24 hours. Overall higher intensity of care occurred earlier and was more likely to be carried out by higher level NNUs. Infants were transferred in way that reflected this, but there are a non-negligible number of infants and transfers which do not meet this pattern. A quarter of transfers were to a different neonatal network. Transferred infants were born at a lower gestational age than non-transferred infants but mortality was lower with fewer deaths within 24 hours and more after 28 days.

Mortality in the cohort was 10%, with around a quarter of deaths occurring within 24 hours of birth, another quarter by 7 days, another quarter by day 28 and another quarter after 28 days. Deaths were more likely to occur in higher level NNUs and surviving infants were more likely to be discharged from lower level NNUs. Thus there are divergent care pathways based on sickness of the infant; these differences were larger for transferred compared with non-transferred infants.

2.4.2 Data source

The NNRD is the only UK data source enabling national, risk-adjusted comparisons of NNU mortality, thus there was no other single data source for the analysis. Perinatal mortality data are specifically collected for MBRRACE-UK (Mothers and Babies: Reducing Risk through Audits and Confidential Enquiries(2016)) but these data only relate

to deaths, with denominator data acquired separately from the ONS. The NNRD contains data on all infants admitted to neonatal care, allowing all providers to be included, and consists of detailed clinical data in contrast to administrative databases. Provider profiling studies often use administrative databases which has attracted criticism for the quality of data (Iezzoni, 2003; Krumholz et al., 2006). While similar concerns exist for routine clinical data, the NNRD is comprised of data captured at the point-of-care, by staff with clinical knowledge rather than coded from medical notes, and some fields are validated as described in section 2.1.3. Regular assessment of clinical outcomes is seen as an essential aspect of quality improvement in health services, but the financial and labour costs of doing so can be prohibitive (Spiegelhalter et al., 2012). As it is derived from electronic patient records, the NNRD has minimal additional costs and is an ongoing collection updated in a timely manner. While the aim of this thesis is primarily to investigate methodology, the suitability of the NNRD for provider profiling purposes is of note. A limitation of the NNRD is that outcomes for infants who are transferred to specialist surgical centres which are not standard neonatal units are not included if they do not return to neonatal care.

2.4.3 Selection of the cohort

The overall cohort consists of all singleton infants born 2011-2013 between 22^{+0} to 31^{+6} weeks gestational age and admitted to a neonatal unit in England, for whom survival to discharge from neonatal care was known. All data analysed in the rest of this thesis are subsets of this cohort. The rationale for these selections was given in section 2.2 but the exclusion of infants with missing outcomes is a limitation of the thesis, potential solutions to which will be discussed as part of further work in chapter 9.

I showed that multiple births are common amongst these infants, though sibling sets cannot always be identified. For twins where matching is possible, their outcomes are similar (i.e. the probability of survival is higher amongst infants whose twin survives). As described in chapter 1 ignoring dependencies violates assumptions of statistical models. For the research questions addressed here there are further implications. Siblings from the same pregnancy are usually treated at the same NNU. Unless correlations within siblings are accounted for in the model, they will be incorrectly assumed to be correlations within NNUs, thus overstating the variation between NNUs.

Correlation of outcomes has been acknowledged and utilised in twin studies for some time (Carlin et al., 2005) yet this correlation is often ignored in studies in the general preterm population (Ananth et al., 2005). In a systematic review of randomised clinical trials in preterm infants, Yelland et al. (2015) found that less than a third accounted for correlation within multiple birth sets in the primary outcome analysis. Investigating and

accounting for correlations within siblings in the NNRD is hampered by the difficulty of matching birth sets. Maternal and infant data which link infants is already used to match siblings, but this is done on a deterministic basis, i.e. a definite match is required. Probabilistic methods such as those used by Croft et al. (2002) might achieve more matches. Even if multiple birth sets can all be identified, these dependencies present particular challenges because the groups (multiple birth sets) are small and unlike twin studies we are analysing all preterm infants (singletons and multiples) so some infants are not part of a multiple birth set. These complexities will be discussed in chapter 9 where I consider possible approaches for extending the methods in this thesis to include multiple births.

Singleton infants are a distinct and clinically relevant population in their own right, but there are some limitations to making this restriction. Almost a quarter of infants were excluded from the analysis because of multiple gestation, reducing power. This could be mitigated by including one randomly selected infant from each pregnancy, but the problems of matching birth sets remains an issue. Multiple births are different from singletons, so excluding these infants may produce different results from an analysis based on the whole cohort. Amongst all births, multiple birth is an established risk factor for perinatal mortality with risks increasing with the number of foetuses (Blondel et al., 2002). However this is not always the case amongst babies born preterm or very low birth weight infants (Slattery and Morrison, 2002), with some studies finding higher mortality for multiple births (Gardner et al., 1995; Tyson et al., 2008), some lower (Doyle and Victorian Infant Collaborative Study Group, 2001; Kiely, 1998) and others finding no difference (Mizrahi et al., 1999; Wolf et al., 1992). As discussed in section 1.1, the mechanisms resulting in early delivery of multiple births are not the same as those for singletons. Furthermore the consequences of preterm birth may also differ. The pattern of relative mortality risk across the gestational age and birth weight range is different for singletons and multiples (Buekens and Wilcox, 1993; Minakami and Sato, 1996). Kiely (1998) showed that while singleton infants born at 36 weeks gestation had a five-fold risk of perinatal death compared with the week corresponding to the lowest mortality (40 weeks), for twins the same relative risk is not reached until 32 weeks (compared with the optimal gestation of 38 weeks). It is also thought that the lungs mature faster in foetuses in a multiple pregnancy than in singletons (Leveno et al., 1984; Ohel et al., 1987), though this has not demonstrably resulted in improved outcomes (Friedman et al., 1997). Therefore inferences about provider effects resulting from analysis of singleton infants alone may not be generalisable to the whole cohort. Furthermore, if we wish to include these infants changes may be required to the risk adjustment, along with a statistical model that accounts for the dependencies in the data.

2.4.4 Infant characteristics

Distribution of infant characteristics in the cohort were compared with ONS data on all live births to illustrate differences between the preterm and term populations. Infants in the cohort were more likely to be male, born by Caesarean section, born to black mothers and come from deprived areas. These findings are in line with risk factors for preterm birth described in chapter 1. Data should be interpreted with caution because levels of missing data were high for some variables (mode of delivery, smoking) and validation against clinical trial data (see 2.1.3) showed that maternal ethnicity was unreliable with 10% of records discrepant. For area-level deprivation, the LSOA and therefore the quintiles are approximately equal in population size so a random individual is equally likely to come from any of the quintiles. However the same is not true of newborns as fertility rates will vary by by LSOA. Nonetheless the distribution illustrates that the burden of preterm birth is borne more heavily by deprived areas.

2.4.5 Structure of neonatal care and transfers

NICUs conduct the majority of care for the very preterm population analysed in this thesis. However Gale et al. (2012b) showed that only 11 NICUs admit more than 100 very low birth weight infants annually which is a recommended criterion for a "high-volume" NNU under centralisation (Phibbs et al., 2007). Although special care constitutes almost all of the care burden of SCBUs, NICUs conduct the majority of special care in terms of volume for the cohort in this thesis. This highlights that infants are not always transferred out to lower level NNUs once they no longer require intensive care. This could be for a number of reasons: the NICU in question is their closest NNU, insufficient cot capacity or transfer facilities, or due to circumstances particular to the infant. Thus neonatal services in the UK are not completely centralised: care is not as concentrated in large NICUs, and NICUs also provide lower levels of care. It has been argued by Phibbs (2012) that a stronger approach closing smaller NICUs would improve neonatal outcomes, as has been seen in Portugal (Neto, 2006). I do not explore the extent of centralisation further as this is not an aim of the thesis, but note that there remains scope for further centralisation which could increase the relevance of this work.

The intentions of the network system and the types of transfer that are required (as outlined in section 1.2.2) are, in general, borne out by the results of sections 2.3.2 and 2.3.3. The pattern of intensity of care with postnatal age, the level of NNU at which care takes place, and the timing and direction of transfers all show a consistent picture. In summary, infants who are not born at a hospital which can accommodate their care needs are transferred, usually within 24 hours, to a higher level NNU for higher level care. Most infants require only one or two transfers. Infants whose condition improves

are transferred back and are discharged from the lower level NNU, while infants that die tend to do so in NICUs. However there are exceptions to this: some intensive care is provided at SCBUs up to day 14, not all transfers on day one are for a higher level of care and there are a non-negligible number of transfers to NNUs of the same level.

Several other studies have examined neonatal transfer patterns. While not an aim of this thesis, the findings from this chapter regarding transfer patterns add to existing literature evaluating networks and transfers (Cusack et al., 2007; Gale et al., 2012b; Redshaw and Hamilton, 2006). The transfer rate was higher than that documented in the Trent region by (Cusack et al., 2007) but similar to the more recent NNRD data reported by Gale et al. (2012b). The frequency of cross-network transfers is documented by the National Neonatal Audit Programme. In 2014, 10% of all infants admitted to neonatal care were transferred at least once, with 17% of transfers being to a different network. The figures reported in this chapter are much higher, reflecting the more complex needs of very preterm infants. I showed that there are differences in the characteristics of transferred and non-transferred infants, with the former constituting over a third of the cohort, so excluding these infants is not desirable. The difference in where deaths and discharges alive occur highlight that the condition of an infant will dictate their location of care.

There is considerable variation in crude mortality but the number of infants at individual NNUs is small. Different patterns are observed when we consider overall mortality by NNU level allocating infants to the final NNU, and allocating infants to all NNUs caring for the infant, but these are quite different measures. Mortality in relation to infant characteristics will be explored further in chapter 3.

The analysis in this chapter will help inform how the structural aspects of the data (level of NNU and transfers) are handled in the provider profiling analysis in subsequent chapters, and aid interpretation of the results. In the next chapter, mortality will be considered in more detail and a risk-adjustment model developed; the inclusion of transfer status and NNU level will also be considered with reference to the findings of this chapter.

Chapter 3

Risk factors for mortality of preterm infants

In chapter 1, I discussed the use of regression analysis as a way of adjusting for risk factors in provider profiling studies. The aim of this chapter is to decide which variables to adjust for and how to model their associations with mortality. These decisions will inform the provider profiling analyses developed in the subsequent chapters. I first describe some considerations for selection of risk factors in provider profiling analyses both in general (section 3.1) and for the thesis (section 3.2). I select infant risk factors based on review of the literature (section 3.3). In section 3.4 the resulting infant factors are examined in the National Neonatal Research Database (NNRD), and a risk adjustment model for infant characteristics is developed and evaluated. Finally in section 3.5 I summarise the findings of the chapter and give the implications for the rest of the thesis.

3.1 Background

In any observational study the association of interest may be exaggerated or attenuated by a *confounder* which is associated with both the predictor and the outcome. For example we may be interested whether birth order affects the likelihood of the child having Down's syndrome. Some of the unadjusted association between birth order and Down's syndrome could be explained by maternal age, which increases with birth order and is a known risk factor for Down's syndrome (Rothman et al., 2008). Confounding can occur in any non-randomised study because the distribution of confounders may not be balanced.

Patients are not randomised to healthcare providers, so we would not expect the distribution of relevant risk factors to be balanced. Some of the imbalance will be the result of differences in the local population (e.g. race and ethnicity, individual and area-level socioeconomic deprivation, employment, education) and other area-based factors (e.g. pollution exposure, quality of other local health services). Commissioning decisions, in particular centralisation, can also lead to differences in the patient population as some hospitals may be limited in the range of services they offer. Risk adjustment (also called *case-mix adjustment*) is a broad topic and the subject of much research (Iezzoni (2003) provides a summary and practical applications). Application of these methods to the mortality of preterm infants is not the primary focus of this thesis, but identification of risk factors is a necessary step before addressing the main research questions.

3.1.1 Risk adjustment for provider profiling

In general we want to adjust for any patient factor which may influence the outcome and is not under the control of the provider (Iezzoni, 1994). Adjusting for these variables reduces any confounding which occurs because a provider has patients that are at higher or lower risk than average through no action of their own. There are a number of considerations to be made when selecting risk factors for adjustment which are particular to provider profiling studies:

Influence of the provider The factors used for risk adjustment should not be those which could be influenced by the provider. In particular, if a complication of care which has a negative impact on the outcome is treated as a case-mix factor, the provider's performance will appear improved after risk adjustment, when in fact it should be lowered (Normand and Shahian, 2007). Conversely if a clinical decision is made which has, say, a positive effect on the outcome via some indicator x, then adjusting for x will not attribute the positive effect to that provider. This is because we are adjusting for something on the causal pathway between our factor of interest (being treated at a particular provider) and the outcome. In a White Paper for the Centers for Medicare and Medicaid Services, Ash et al. (2012) recommended, based on consensus from all stakeholders involved, that only risk factors present at admission are considered to avoid conflation with provider effects.

Data quality and other systematic differences It has been shown that estimates of risk-adjusted outcomes are sensitive to differences in coding (Bottle et al., 2011; Mohammed et al., 2009). Similar biases will occur in any situation where the recorded value for the same patient would differ systematically by provider. Reasons for this include

measurement differences in equipment or technique, or subjective measures which may vary by assessor. Concerns over measurement differences can be minimised by selecting variables which are clearly defined and accurately recorded (Normand and Shahian, 2007; Spiegelhalter et al., 2012). Data completeness is also important, because if missing outcomes differ to those observed, results will be misleading, particularly if patterns of missing data vary across providers (section 2.2.7 includes a discussion on types of missing data).

Sociodemographic data Sociodemographic variables such as ethnicity and deprivation, and variables related to health behaviour such as parental smoking and alcohol consumption, are often considered as candidate risk factors in epidemiological studies for a range of outcomes (Iezzoni, 2003). These variables usually have some spatial variability and may therefore confound the provider effect on the outcome, as healthcare providers predominantly serve the local community. As with clinical risk factors, this can lead to some providers having better or worse outcomes due to the sociodemographic profile of the local population and could be considered as factors to adjust for. Several authors have argued against adjusting for sociodemographic risk factors in provider profiling studies for a number of reasons. If sociodemographic variables are correlated with the provider effects themselves (e.g if true poor performers are located in more deprived areas) then controlling for such variables may adjust for some of the factors we wish to examine (Normand and Shahian, 2007). The correlation may not just be induced by geography but could be intrinsically linked with provider performance. For example Normand and Shahian (2007) argue that ethnicity and socioeconomic status should not be included in risk-adjustment because they may be associated with system-level factors (e.g. lack of translators or poor community outreach services). Thus adjusting for sociodemographic factors mask situations where poor care is being provided to disadvantaged patients, and could lead to providers believing that a lower standard of care for these patients is acceptable (National Quality Forum, 2014). Sociodemographic variables are likely to be of inconsistent magnitude across the country because they are often proxies for underlying risk factors with direct physiological effects (e.g. poor diet, stress) (Nicholl, 2007). Applying a uniform risk adjustment across all providers is therefore not reasonable, and while we can model varying effects across providers (see section 4.1) the inability to distinguish genuine variation in risk factor effects from provider-induced variation remains. Rather than include sociodemographic variables in risk adjustment, it has previously been recommended that such data be used to give context to any differences that emerge (Ash et al., 2012; Iezzoni, 2003; Krumholz et al., 2006; Normand and Shahian, 2007).

The issue was subsequently re-examined by an expert panel convened by National Quality Forum (NQF), a US organisation that promotes quality improvement in health-

care (National Quality Forum, 2014). The driver for the re-examination was a concern about the fairness in providers being held to account on the basis of outcomes which were not adjusted for sociodemographic factors, particularly with regard to payment-by-results and sanctions for poor performance. The NQF panel suggested that this could also lead to further inequity in healthcare through a number of mechanisms. If providers serving disadvantaged communities are financially penalized for their results, subsequent care may suffer. Funding will also be reduced for such providers if their results lead to patients choosing to go elsewhere. Patients who are high-risk based on their sociodemographic profile may have reduced access to care, as decisions on patient catchment areas or expansion of facilities may be made to optimise reported outcomes and any associated financial reward. The conclusion of the NQF panel was that on balance these concerns outweighed the existing arguments against adjusting for sociodemographic factors. They recommended that provided there was conceptual and empirical support for the an association between a sociodemographic variable and the outcome, and the variable met the same criteria applied to clinical risk factors, then it should be included in the adjustment, but that the measure should be presented adjusted for clinical factors only, stratified by the sociodemographic variable.

Provider level variables As with patient-level risk factors we do not want to adjust for any provider-level variables which are caused by the provider, for example a quality improvement programme initiated internally. Provider-level factors could be associated with the outcome because they encapsulate unmeasured case-mix differences between provider populations (urban or rural location), because they affect the ability of the provider (the number of beds available) or both (whether a hospital is public or private). Provider-level summaries of individual-level characteristics can be included to yield more nuanced inferences about individual-level effects. For example, Aitkin and Longford (1986), when investigating variation in examination results across schools, included an IQ-type measure both at the pupil level and aggregated to school level to investigate whether pupils do better when they are at school with either high or low scoring pupils.

Some provider-level factors might be considered noise if the aim of the analysis is to distinguish variation which is not just associated with each provider, but beyond that attributable to its characteristics. Austin et al. (2004) found that including hospital peer group (an indicator grouping hospitals based on patient volume) accounted for more between-hospital variation and resulted in fewer outlying hospitals than an analysis using patient characteristics alone. It allowed estimation of performance relative to the peer group average, but also of the average differences between peer groups. Therefore including provider-level variables has another purpose in attributing variation between providers to different sources.

The decision to adjust for a provider characteristic is not always clear, as there are some which could be viewed as attributes which the provider should be credited for. Ash et al. (2012) give the example that if a hospital has a special cardiac care unit, their result should incorporate this benefit rather than be conditional on it. If the presence of such a unit is not included in the risk adjustment model the average effects across all providers is not estimated, but this may not be a suitable aim for some attributes.

It is often the whole profile for a provider, combining different interacting characteristics, that influences the outcome in a complex manner rather than individual variables acting independently, in which case separate examination of individual providers may be more appropriate. In the review of NHS Trusts undertaken by Sir Bruce Keogh, the Medical Director for NHS England (Keogh, 2013), fourteen trusts with unusual performance were identified based on measures adjusted for patient risk factors only. A subsequent detailed assessment of performance was made in the light of other provider-level data such as staff levels and sociodemographic data for the area. Similarly, the Summary Hospital Mortality Indicators (SHMI) for English hospitals are published alongside a series of contextual indicators to help users interpret the results (Health and Social Care Information Centre, 2016b). One such indicator is palliative care, which is not adjusted for because of the variability in how data are recorded. Figure 3.1 shows an example from the 2015 report (Health and Social Care Information Centre, 2016b) illustrating where NHS Trusts who are repeat outliers for mortality lie regarding the level of palliative care coding. Of the ten trusts whose mortality was lower than expected, seven had a higher than average proportion of deaths recorded as palliative care. If palliative care had been adjusted for, the SHMI for these trust would have been even lower.

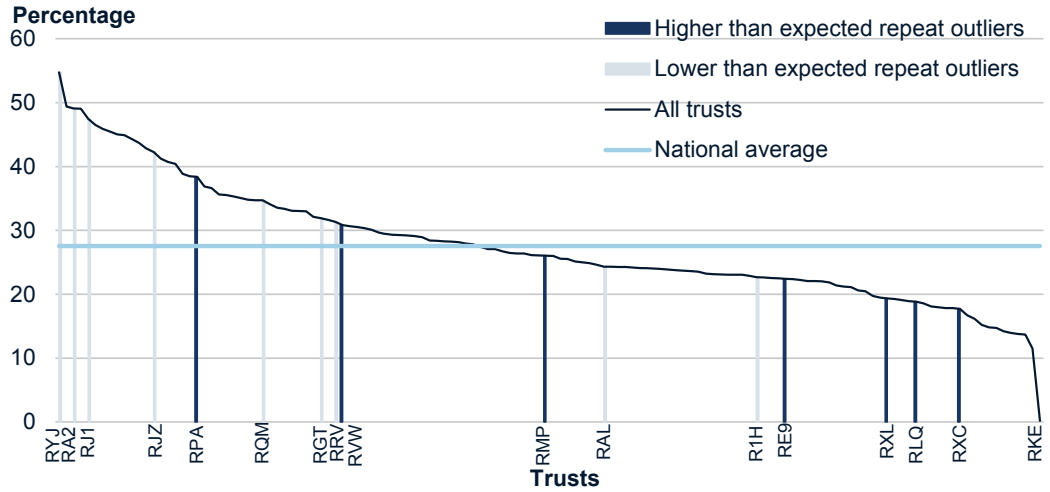


Figure 3.1: Percentage of deaths with palliative care coding for all NHS Trusts and those which are repeat outliers, January 2015-December 2015; © 2015 Re-used with the permission of the Health and Social Care Information Centre. All rights reserved.

3.2 Risk adjustment for the thesis

Based on the above considerations, at the individual level I will only adjust for established risk factors measured at or before birth which are not influenced by neonatal care. I will consider several factors at the provider level, and again these should not be influenced by the provider. For feasibility with regard to implementing the methods developed in this thesis as part of routine reporting, all factors should be available in the NNRD, clearly defined and well recorded.

3.2.1 Clinical risk factors

Externally developed risk scores are often used for case-mix adjustment in provider profiling studies (DeLong et al., 1997; Glance et al., 2003). They have the advantage of being externally validated and reducing risk of subjectivity but may not capture illness severity well for the cohort in question. Several risk scores have been developed for the neonatal population such as CRIB (Clinical Risk Index for Babies) II (Parry et al., 2003), SNAPPE (Score for Neonatal Acute Physiology Perinatal Extension) II (Richardson et al., 2001) and PREM (Prematurity Risk Evaluation Measure) (Cole et al., 2010). These scores have been developed to aid decision making for clinicians and parents as well as for risk-adjustment. Some of these scores include variables which might reflect care received. For example, CRIB II (Parry et al., 2003) includes temperature at admission which is is a marker of the condition of the baby but can be influenced by care (Guenther et al., 2015; Manktelow et al., 2010). Others include variables not well recorded in the NNRD, such as base deficit (a measure of the acid-base balance in the blood) which is part of CRIB II and the PREM score (Cole et al., 2010). Some scores are only suitable for a subset of the infants included in the cohort here. For example the Neonatal Research Network of the National Institute of Child Health and Human Development developed a prediction model for infants up born 22 to 25 weeks gestation (Tyson et al., 2008). Given the multitude of models available and the inclusion of unsuitable or unavailable variables in some models, I chose to develop an internal risk-adjustment model, basing choice of risk factors on those used in existing models to ensure only established predictors were included and to reduce subjectivity in variable selection (Glance et al., 2006). This will be described in sections 3.3 and 3.4.

3.2.2 Sociodemographic factors

Some studies of variation perinatal mortality such as the reports produced by MBRRACE-UK (Manktelow et al., 2016) have included socio-economic status (SES) as a risk factor.

Associations between SES and both increased infant mortality and increased risk of preterm delivery have been documented (Gray and McCormick, 2009; Kramer et al., 2001) but the influence of SES on mortality in the preterm population is less well established (Smith et al., 2009). The relationship is difficult to assess partly because of the number of contributing factors (for example income, education, employment, with measures possibly considered at maternal, paternal and neighbourhood levels). Calling et al. (2011) found substantial increases in infant mortality comparing low and high income families in a study of nearly 50,000 preterm infants (< 37 weeks) in Sweden. Smith et al. (2009) found no evidence of a link in the Trent region of the UK, though the study was smaller, the population different (< 32 week infants) and area-level rather than individual measures of deprivation were used. A fundamental issue for assessing the influence of SES on mortality for preterm infants is that the causal relationships are complex: preterm birth is on the causal pathway between low SES and increased mortality, so conditioning on gestation by only examining preterm infants may result in biased inferences (Hernán et al. (2002) illustrate this phenomenon in a perinatal context). Neither Calling et al. (2011) nor Smith et al. (2009) raised the difficulty of drawing causal inferences when conditioning on a mediating factor.

In the NNRD deprivation is only available at the area level rather so may not be an accurate reflection of the circumstances at the individual level which is of interest here. In particular for provider profiling, some areas may be more mixed than others regarding deprivation, so the suitability of the area-level measure may be inconsistent across providers. The only individual-level measure of SES in the NNRD is parental occupation, but this is free text and is poorly completed. Ethnicity exhibits some of the same risk-adjustment issues as SES: racial and ethnic groups have varying rates preterm birth and infant mortality (Goldenberg et al., 2008; Manktelow et al., 2016), and between-provider variation is likely to be due to differences in the local population. Maternal ethnicity is over 90% complete but some categories are ambiguously defined (see table 2.2). Paternal ethnicity is only 70% complete and ethnicity of the infant is not recorded in the NNRD.

In section 3.1 I discussed the advantages and disadvantages of adjusting for sociodemographic factors. Neonatal units (NNUs) are not rewarded or sanctioned on the basis of their mortality rates, nor is this information currently publicly available, although mortality data are published by neonatal network. Therefore some of the concerns that lack of adjustment will drive inequality are not as applicable here. Given the quality of SES and ethnicity data, and the conceptual difficulties of the associations with mortality, I have elected not to adjust for sociodemographic factors in this thesis. This has implications for the interpretation and utility of the results which will be discussed in section 3.5. Stratification is not reasonable due to low number of infants at each NNU, but area-level deprivation and maternal ethnicity are considered as contextual factors

in the case study of chapter 8.

3.2.3 Adjusting for transfer status

In chapter 1 I described how transfers are an integral part of networked neonatal care because not all services are provided at every NNU. Transfers in the NNRD were analysed in section 2.3.3, which showed different patterns in infant characteristics, mortality, timing of deaths and location of care. The need for transfer indicates that the infant has complex needs, but conversely that they have survived long enough and are stable enough to be transferred. This was demonstrated in chapter 2 as transferred infants had lower gestational age and birth weight but there were fewer deaths within the first 24 hours. In addition to the findings in the previous chapter, several studies have demonstrated increased mortality or morbidity following transfers (Delaney-Black et al., 1989; Hohlagschwandtner et al., 2001; Modanlou et al., 1979; Towers et al., 2000).

Therefore transfer is associated with the mortality, and as demonstrated in section 2.3.3 transfer rates vary across NNUs. This raises the question of whether we should control for transfers in a provider profiling analysis. The need for transfer may tell us further information about the condition of the infant which is not captured by the known risk factors. However transfers are part of care provided by the NNU currently caring for the baby as it is their responsibility to arrange the transfer, and so is on the causal pathway between the provider's actions and the outcome and should not be adjusted for (section 3.1.1). Appropriate transfers are a sign of good care, but if the need for transfer arose after birth it is possible that this need was a result of inadequate care. The need for postnatal transfer could be a result of the failure to organise an in-utero transfer (Gale et al., 2012a). Whether the transfer is part of care from the point of view of the receiving NNU is less clear. While they have not yet cared for the infant, they are involved in the arrangement of the transfer, and failed in-utero transfer may be partly attributable to the receiving NNU (e.g. because of capacity management issues or communication problems) (Gale et al., 2012a; Kempley et al., 2006). Network policies and attributes are likely to have an impact on whether a transfer occurs, and when it occurs. While there are national guidelines, networks have their own care pathways contingent on the properties and locations of the NNUs in the network, and these specify which infants should be transferred to a NICU. Initiatives such as a dedicated transfer service to help organise in-utero transfers will affect the likelihood and timing of transfer (Kempley et al., 2006). Therefore from a network perspective transfers are under the influence of care.

Within the data available in the NNRD, reasons for transfer do not meet the criteria of being well-defined clinical data (section 2.3.3). Whether an infant was transferred in-

utero cannot be determined from the NNRD. We do know if a mother booked to deliver at a different location to the first neonatal admission, but we do not know whether she presented for delivery there and was subsequently transferred. The need for transfer is merely a surrogate marker for underlying conditions, and the data we have only indicate postnatal transfers which actually occurred. Different types of transfer will indicate different illness severity but these types are not clearly defined and are not estimable from the data. We also do not know whether the need for transfer was present at birth. This cannot be assumed even for very early transfers as even the first minutes and hours after birth are crucial (Empana et al., 2003; Fenton et al., 2004; Mohamed et al., 2010).

Another factor to consider in risk adjustment is whether the variable has a consistent effect across providers. Non-transferred infants include successful in-utero transfers as well as those not needing transfer. If, on average, transferred infants have poorer outcomes, then adjusting for transfer will penalise NNUs and networks who are better at facilitating in-utero transfers as these infants will be assumed to have the risk level of a non-transferred infant. As previously mentioned the need for transfer might be induced by inadequate care; in such cases controlling for transfer will "adjust away" the very effect we are interested in.

Therefore I am not going to adjust for transfers as a risk factor in this thesis but will consider results in light of transfer rates in chapter 7.

3.2.4 Provider-level variables

Neonatal units are designated at different levels depending on the types of care they provide (see section 1.2) so we would expect there to be variation in case-mix across NNU levels, some of which is accounted for by known, measured risk factors. It may also be that NNU level captures unmodelled variation in case-mix as additional (possibly unmeasured) information about the sickness of the infant available before birth may determine where they are cared for, either via planned place of delivery or in-utero transfer. In section 2.3 I examined how the volume and intensity of care varies across NNUs. Even within this cohort of very preterm infants SCBUs provided a lot less intensive care, and a much lower volume of care. Deaths are more likely to occur in higher level NNUs, and discharges home (or to a paediatric ward) are more likely to occur from a lower level NNU. Therefore it may not be sensible to compare across NNUs levels. NNU level is also a an objective, clearly defined and pre-assigned measure. One of the aims of case-mix adjustment is to compare like with like; including NNU level in the model allows us to do this, and to estimate any difference between levels as was discussed in section 3.1.1.

Volume of care is another NNU characteristic which is sometimes adjusted for in provider profiling studies, including in neonatal care (MacNab et al., 2005). It can be specified in a number of different ways, for example the number of very preterm admissions, or the number of intensive care days. Adjustment for volume to reflect peer group may not be necessary in the neonatal case as NNU level already serves this purpose, however the analysis in section 2.3.2 shows there is considerable variation in volumes within NNU level. The association between higher volumes and improved outcomes forms the basis of the clinical arguments for centralisation (Cifuentes et al., 2002; Phibbs et al., 1996). Watson et al. (2014), in an instrumental variable analysis of NNRD data, showed a negative causal effect of case volume on mortality. However, this does not preclude a causal association in the opposite direction: well-performing NNUs may through reputation attract higher volumes. In the UK, mothers can book to deliver wherever they choose, though this requires a choice of hospitals offering maternity services within a reasonable distance. If this is the case, volume is on the causal pathway between NNU performance and mortality, so I have chosen not to control for volume in this thesis. In an approach similar to reports for the SHMI depicted in figure 3.1, I will illustrate how descriptive data on NNU volume can contribute to a case study for a single NNU in chapter 8.

The UK Neonatal Staffing study (2002) found that NICU mortality increased when operating at either higher capacity or with higher infant-to-nurse ratios. Cot occupancy, staffing levels and a perhaps summary measures of the characteristics of co-patients could therefore be considered as risk factors. As discussed in section 2.3.3 it is not possible to determine capacity or staffing levels from NNRD data. It is also arguable that capacity should be considered part of the NNU effects, and not a factor beyond provider control which should be included in the adjustment. Therefore I do not consider capacity further here. Other characteristics which are not attributes of the NNU but could be considered as provider-level variables are economic indicators for the area, such as unemployment rates. I am not considering adjusting for the reasons described in 3.1.1, but will include area-level deprivation as a factor in the case study in chapter 8, along with maternal age and ethnicity, to illustrate how such data can give context to mortality indicators.

The statistical methods for including provider-level variables will be reviewed in chapter 4, and implementation set out in chapter 5. For the remainder of this chapter I focus on developing a model for individual-level risk factors.

3.3 Selection of infant risk factors

3.3.1 Literature review

Medlock et al. (2011) carried out a systematic review of prediction models for mortality of preterm infants, one of the aims of which was to identify important predictor variables. I selected infant risk factors based on the findings of this review. The authors searched the MedLine database for studies which provided mortality predictions for infants born before 32 weeks gestation and/or below 1500g birth weight, or subpopulations of this group provided they did not have a specific disease or condition. Some assessment of model performance had to be reported for the study to be included. Full methods for the study and the resulting papers can be found in the published paper.

As there may have been further developments since publication of the study, I updated the search using the same search strategy for papers published after May 2010 (the search date given by Medlock et al.). The search is up to date as of 11th April 2016. This yielded 25 further relevant studies but this did not yield any further variables (see appendix B for details). Two other reviews of predictive models were identified (Dorling et al., 2005; Pollack et al., 2000), neither of which were systematic reviews. These studies were identified by Medlock et al., and so did not yield any additional models.

There were 41 studies included and 11 variables which the authors deemed to be frequent predictors: gestational age (included in 24 studies out of 28 which examined the variable), birth weight(29/36), appropriate size for gestational age (AGA, defined as between the 10th and 90th percentiles of birth weight for gestation, 10/22 studies), sex (18/30), ethnicity (9/20), antenatal steroids (11/20), Apgar score(18/26), temperature on admission (5/12), measures of respiratory status (blood gas (11/16) and FiO_2 (6/7) and respiratory function (9/19)).

Of the variables listed above, ethnicity, Apgar score, admission temperature and the three respiratory measures were not considered suitable risk factors for provider profiling purposes. I have chosen not to adjust for ethnicity for reasons described in 3.2.2. The Apgar score is a quick assessment of newborn health by carried out in the first minutes after birth, but is a subjective measure with high interobserver variability (O'Donnell et al., 2006). It is therefore possible that Apgar scores could be inconsistent across NNUs and adjustment could bias estimates of NNU mortality rates via mechanisms described in 3.1.1. If staff members at the same NNU are likely to give similar Apgar scores, this bias could be even stronger. Temperature and respiratory measures on admission were excluded on the basis that they can be affected by neonatal care, particularly as NNU staff are often present at the delivery of very preterm infants (Guenther et al., 2015; Manktelow et al., 2010). Steps can be taken in the delivery room to raise the

temperature of an infant prior to admission (e.g.occlusive wraps and radiant warmers (Royal College of Paediatrics and Child Health, 2015)), and decisions on respiratory support are usually made in the delivery room (Guenther et al., 2015).

This left gestational, birth weight, size for gestation, sex and antenatal steroids as risk factors for inclusion in the model. I shall describe these factors in general terms in section 3.3.2 before descriptive analysis and modelling using NNRD data in section 3.4.

3.3.2 Description of risk factors and patterns of association

Gestational age Gestational age estimates can differ depending on whether they are determined from the date of last menstrual period (LMP), or by the ultrasound dating scan offered to all women at week 10-14 of pregnancy. The NNRD contains the *best obstetric estimate* which is initially based on LMP and can be modified on the basis of antenatal ultrasound. This estimate of gestational age is unaffected by the NNU since both the LMP and scan estimates would have been derived during antenatal care. If neither LMP nor scan data are available at delivery then the gestational age recorded is based on a postnatal estimate of maturity by clinical staff. There may be differences in the postnatal estimates of maturity, but since some estimate of gestation should always be made as part of routine antenatal care (National Institute for Health and Care Excellence, 2008) it is unlikely that a postnatal estimate would be required.

The relationship between gestation and neonatal mortality (within 28 days) across the gestational age range was illustrated in figure 1.2 using data from all livebirths in England and Wales from the Office of National Statistics. This showed that the risk of mortality is not uniform within the preterm population. Most of the variability occurs in the gestational range below 32 weeks, which is the population for this thesis. Of the livebirths in England and Wales from 2007 to 2013, 93% of those born at or before 22 weeks died within 28 days. Mortality decreases dramatically with each additional week of gestation for these extremely preterm infants: 69% died after birth at 23 weeks, 38% at 24 weeks, 23% at 25 weeks, 15% at 26 weeks and 10% at 27 weeks. After this the reductions are less steep, with neonatal mortality of infants born at 31 weeks being 2.2%. Thus the benefit of an additional week in-utero is non-linear, decreasing as gestation increases, though as previously described in section 1.1 a survival disadvantage remains for those born even slightly before term.

Birth weight As discussed in section 2.2.2, birth weight is often used to classify infants, with low birth weight indicating those at higher risk of mortality and morbidity, even through to later life. I have chosen to use gestational age to define the population analysed here, but within preterm infants birth weight is still an important predictor,

as demonstrated by the frequency with which it appears in the prediction models for survival of preterm infants described above.

Higher birth weight is not always better: when mortality across the birth weight range for is examined for all newborns we see a reverse J-shaped pattern consistent across populations (Basso and Wilcox, 2011; Wilcox, 2001). This shape is a steep decline in mortality with birth weight for much of the range with the lowest rates lying between one and two standard deviations above the mean, followed by an uptick for the highest values. This pattern of adverse outcomes at the relatively higher birth weight values is not so established for preterm infants, but has been demonstrated in some studies. Da Frè et al. (2015) found higher odds of mortality for those on or above the 90th centile compared with those in the 75th to 89th centiles amongst infants born less than 32 weeks . In developing the PREM score for predicting survival for infants born less than 32 weeks gestation, Cole et al. (2010) found a notable survival disadvantage in infants with birth weight two standard deviations above the gestation-specific mean, with the strength of the disadvantage increasing with gestation. Models were fitted in two separate datasets and the non-linearity was observed in both, reducing chances of a spurious association. Cooke (2007) found that for infants born below 35 weeks gestation, there was no difference in mortality for infants with a large birth weight for gestation (greater than three standard deviations above the mean) compared with those of average size, though the length of the study (25 years) could confound the associations as both overall survival and composition of the population is likely to have changed in the period.

Size for gestation As previously described birth weight can be expressed relative to gestational age using z-scores (or equivalently percentiles) which indicate where the infant's birth weight lies on the gestation-specific distribution. The percentile of birth weight for gestation is used to determine whether an infant is small (less than 10%), appropriate (10 to 90%) or large (greater than 90%). It has been argued that these cut-offs are arbitrary as associations between fetal growth and subsequent health outcomes are continuous (Oken et al., 2003; Wilcox, 2001).Such definitions may be useful for identifying a cohort or for comparing rates of growth restriction in groups, but is an unnecessary simplification for modelling purposes where continuous data are available as is the case here.

Sex Males are at higher risk of many adverse health outcomes throughout life, and the perinatal period is no exception (Office of National Statistics, 2008; Peacock et al., 2012). Preterm birth is more common in boys (Ingemarsson, 2003; Zeitlin, 2002); following preterm birth boys are more likely to die in the neonatal period, due in part to lagging

lung maturity (Costeloe et al., 2000; Peacock et al., 2012).

Antenatal steroids As discussed in section 1.1.3 routine administration of antenatal steroids to women in preterm labour has played a part in the dramatic increase of survival rates of preterm infants by improving lung function. The current guidelines from the Royal College of Obstetricians and Gynaecologists (2010) state that the steroids are most effective when delivery occurs between 24 hours and 7 days after the second dose is given, so preterm birth should ideally be anticipated at least 24 hours in advance of delivery. In the NNRD, administration of antenatal steroids is recorded in two fields. One records whether any dose of steroid was given at any point during pregnancy with the intention of improving fetal lung maturation and the other records whether a complete or incomplete course was given, though this does not include the timing and completeness was poor for this field.

3.4 Statistical modelling of infant risk factors

In this section I present NNRD data for the associations with mortality for each of the risk factors selected in the previous section (gestational age, birth weight, average size for gestation, sex and antenatal steroids). I then combine the risk factors in a multivariable regression model and assess its predictive performance. Other methods of standardisation exist which do not involve regression, but these are generally only suitable for categorical risk factors (Aylin and Bottle, 2007). Note that the main methods for the thesis will be presented in chapter 5; what follows here is a preliminary analysis to establish the functional form of the continuous risk factors and assess the suitability of the risk adjustment model prior to implementing hierarchical methods for the provider profiling analysis. Hence I have not used hierarchical models here and the standard errors for the covariates may be underestimated at this stage (see section 4.1 for further discussion of this).

3.4.1 Methods

The data analysed in this section were as described in section 2.3. All statistical analysis was carried out in Stata version 12 (StataCorp., 2011). Associations with mortality for each risk factor were first described using tables. Death before discharge was modelled using logistic regression. All variables selected in section 3.3 were kept in the model; the purpose of model selection here was to determine the shape of the associations with continuous variables and how to model interactions. Let p_i denote the probability that infant i dies before discharge. To model the association between mortality and variables

$X_1, ..., X_m$ which take values $x_{i1}, ..., x_{im}$ for infant i then we have:

$$logit(p_i) = log(\frac{p_i}{1-p_i}) = \alpha + \beta_1 x_{i1} + ... + \beta_m x_{im}$$

Whether the mother received antenatal steroids was considered as a binary variable for any dose, as data completeness was poor for whether a full course was given, and this was how steroid use was generally treated in other prediction models in the review by Medlock et al. (2011). For binary variables (sex of the infant and whether the mother received antenatal steroids), exponentiating the β gives the estimated odds ratio (OR) comparing one group the other. Continuous variables (gestational age and birth weight) were mean-centred and standardised, so e^{β} gives the OR for a one standard deviation increase in X, assuming a linear relationship between $logit(p)$ and X.

Functional form for continuous variables

Linearity for continuous variables was checked by plotting the logit of observed mortality rates against categories of X. Non-linearity can be modelled using polynomials but these can have undesirable properties (see section 3.5.3). Cubic plines are polynomials in X joined at "knots" which allow the functional form to vary across the range of X. Forcing the first and second derivative of the functions either side of the knot to agree smooths the spline at the knots. The tails (before the first knot and after the last), are also polynomials, but this can yield a poor fit for extreme observations (Harrell, 2001). I therefore used *restricted cubic splines* (RCS), also called *natural splines*, which are like cubic splines but are 'restricted' because they are constrained to be linear in the tails to give improved fit. From Harrell (2001), to model a variable X with a K-knot RCS, with knots at values $k_i, i = 1, ..., K-1$ the terms $X_1, ..., X_K$ to be included in the regression model are given by:

$$X_1 = X_0$$
$$X_{i+1} = \frac{(X_0 - k_i)_+^3 - (k_K - k_{K-1})^{-1}\{(X_0 - k_{K-1})_+^3(k_K - k_i) - (X_0 - k_K)_+^3(k_{K-1} - k_i)\}}{(k_K - k_1)^2}$$
$$i = 1, ..., K-2 \quad (3.1)$$

where

$$(a)_+ = a \qquad a > 0$$
$$0 \qquad a \leqslant 0$$

Models are in general sensitive to the number of knots but not to the location of the knots (Durrleman and Simon, 1989; Harrell, 2001). I selected the number of knots from three,

four or five based in minimising the Akaike Information Criterion (AIC) (Akaike, 1974) and by inspecting plots of predicted values. The AIC is a measure of goodness-of-fit, penalised for the number of parameters so as to favour more parsimonious models and avoid overfitting. The knots were at equally spaced percentiles of the variable, between the 10th and 90th percentiles fo three knots and between the 5th and 95th percentiles for four or five knots (Harrell, 2001).

Functional form for interactions

I followed recommendations of Harrell (2001) regarding model building with RCS. I applied a Wald test to the joint null hypothesis that all the spline terms are equal to zero, and if the null hypothesis was not rejected the variable was modelled as linear. Then I plotted the spline function against the linear variable to check whether a the spline could be replaced with a simpler function. If this appeared the case, I compared the AIC of the spline model with that of the simpler function. If similar, I checked whether there was evidence of non-linearity beyond the simpler function by creating a 5-knot spline of the simpler function and testing the joint hypothesis that all these terms are zero. The aim of replacing the spline with a simpler function was to build a more parsimonious model.

As described in section 3.3.2, gestational age and birth weight have interacting influences which is one reason why z-scores are used. Reference populations used to calculate z-scores differ, although there is one standard in the UK endorsed by the Royal College of Paediatrics and Child Health (Wright et al., 2010) which has been extended to include infants born 23 weeks or later. I modelled birth weight in grammes rather than using the standard z-score because it allows infants born at 22 weeks to be included.

I modelled the weight for gestation relationship using interaction terms in the regression. This also avoids loss of information through categorisation of infants based on size for gestational age as was discussed in 3.3.2. To determine any interactions involving spline terms, I began with the full model of all cross-products of the gestation and birth weight terms. If the gestational age terms are denoted $G_1, ..., G_S$ and the birth weight terms are denoted $B_1, ..., B_T$ then there are ST possible interactions given by $G_1B_1, ..., G_SB_1, G_1B_2, ..., G_SB_2, ..., G_1B_T, ..., G_SB_T$. The strategy I took to reduce the model was to apply Wald tests to check groups of interactions, testing the joint null hypothesis that all terms are zero (Harrell, 2001):

- Full model vs. no interaction (test of ST terms)
- Full model vs linear x linear interaction G_1B_1 (test of $ST - 1$ terms)

- Full model vs interactions involving linear birth weight terms $G_1B_1, ..., G_SB_1$ (test of $ST - S$ terms)
- Full model vs interactions involving linear gestation terms $G_1B_1, ..., G_1B_T$ (test of $ST - T$ terms)
- Full model vs interactions involving linear terms $G_1B_1, ..., G_1B_T, B_1G_2, ..., B_1G_S$ (test of $ST - S - T + 1$ terms)

I reduced the model by removing the smallest set of term for which the null hypothesis was rejected based on the tests above, using a p-value threshold of 0.05. I then repeated the procedure to compare the resulting model with simpler models.

Plots of standardised Pearson residuals and Pregibon leverage were used to check for poor fit, departures from linearity and influential observations. The residual is a measure of how much the observed value deviates from the predicted value (exact definition can be found in Hosmer et al. (2013)). The leverage is a measure of how much an observed covariate pattern deviates from average covariate patterns (see Pregibon (1981) for definition).

Model performance and model comparison Discrimination (ability to differentiate between infants that survived and those that died) was checked by calculating the *concordance-statistic*, or *c-statistic* (Harrell et al., 1996). This is defined as the proportion of all possible pairs of individuals with discordant outcomes (where one has the event and one does not) for which the model assigns the higher probability to the correct individual. The c-statistic is equivalent to the area under the receiver operating characteristic curve.

The Brier score was calculated as a measure of overall model fit (Brier, 1950). The Brier score is the mean squared error of the prediction. If infants $i = 1, ..., N$ have observed outcomes $y_1, ..., y_N$ where 1 indicates death before discharge, the Brier score is given by:

$$\sum_{i=1}^{N}(y_i - \hat{p_i})^2$$

Thus the Brier score takes values from 0 to 1 with smaller values indicating better fit.

Calibration (comparability of actual and predicted survival) was checked using Cox's calibration (Cox, 1958; Miller et al., 1991). A logistic regression is constructed for the outcome of interest against the linear prediction from the model that is being checked (i.e. the logit of the predicted probabilities). If the model predicts perfectly the regression intercept will equal 0 and the regression slope will equal 1. The intercept

indicates if the overall predictions are too low or too high, with positive values indicating under-prediction. The regression slope indicates if the predictions are under or over-dispersed, with values over 1 indicating under-dispersion. By definition perfect calibration is achieved when the analysing the whole dataset used to build the model, so I applied the calibration to gestational age subgroups (less than 28 weeks, and 28 weeks or above).

Model performance may be over-estimated as the performance was assessed on the same dataset used to build the model. Therefore the *c*-statistic, Brier score and Cox's calibration measures were corrected for optimism using bootstrapping (Harrell, 2001). I took a bootstrap sample (a sample with replacement) of the dataset and estimated the regression model on the sample with the same covariates but resulting in different parameter estimates, denoted $model_{boot}$. Each performance statistic for $model_{boot}$ was calculated on the bootstrap sample and the original data and $optimism_{boot}$ taken as the absolute difference between these. This process was repeated 2000 times and an estimate of the overall optimism was obtained as the average value of $optimism_{boot}$.

The performance of the model was compared with three existing models to predict death before discharge amongst preterm babies admitted to neonatal units:

1. CRIB II (Parry et al., 2003), a frequently used score based on data from infants born in 1998-9 and admitted to a random sample of 35 UK NICUs;
2. a more recent UK model using data from The Neonatal Survey (TNS), an ongoing data collection from NNUs in the East Midlands and Yorkshire region; (Manktelow et al., 2013)
3. the National Institutes of Child Health and Development (NICHD) model (Tyson et al., 2008) based on infants born 22^{+0}-25^{+6} weeks gestation in 1998-2003 admitted to 19 US hospitals.

While the NICHD model is based on older data from a different country for a different gestational age range I have included it as a comparator because the accompanying web-based prediction tool (National Institute for Child Health and Human Development) means it is often used in clinical practice. Where the comparator model was restricted to certain neonatal characteristics, the relevant subset of the NNRD cohort was used to match the population. CRIB II required base deficit data which was only available in 40% of observations and the TNS model is for white infants born at 23 weeks gestation or over. The predicted survival rate, *c*-statistic, Brier score and the intercept and slope from Cox's calibration were compared between the NNRD and comparator models. For Cox's calibration the data were not split by gestational age as the data were already subsetted according to the relevant population of each model.

Non-linear terms and interactions mean the odds ratios for gestational age and birth weight are not easily interpretable, therefore I constructed charts similar to those developed for the PREM score (Cole et al., 2010) to show associations in a more meaningful manner. Contours indicating infants with the same survival probabilities are shown across the gestational age and birth weight range. Birth weight relative to gestation is shown by the birth weight centiles, calculated from the Royal College of Paediatrics and Child Health (RCPCH) growth charts (Wright et al., 2010) which extend to preterm infants. The probability of survival rather than death was shown as these charts were intended for use in clinical practice.

3.4.2 Results

The overall distribution of all risk factors was shown in table 2.2. Associations with mortality are shown in table 3.1. Mortality decreased with increasing gestational age. It also decreased with increasing birth weight with a slight increase in mortality for the highest birth weights, though the numbers were small. For gestational age and birth weight the associations are both non-linear, as shown in figure 3.2 which plots $logit(p)$ for the observed probability of death in each category of gestation and birth weight in table 3.1. Mortality was higher for males (10.9% vs 9% for females) and for those whose mothers were not given steroids antenatally (18.8% vs 8.8%). The only risk factor with any missing data was antenatal steroids, which was missing for 217 (1.3%) infants. The mortality rate for those with missing data was 10%, identical to the cohort as a whole (see table 2.2). These infants were excluded from subsequent analysis as numbers were small. Note that the infants with missing birth weight z-score are those born at 22 weeks gestation to whom the RCPCH reference does not apply.

Functional form for continuous variables

Table 3.2 shows the AIC for models for gestational age and birth weight using 3, 4 and 5 knots.

For gestational age the 3-knot model was selected for parsimony; figure 3.3(a) shows that there is little difference in the predicted probabilities, though the 5-knot model predicted slightly higher mortality for infants born at 22 weeks, of which there are few. A 3-knot model has 2 terms, one of which is linear. The second term improved the fit of the model ($\chi^2 = 19.7$ on 1 degree of freedom (d.f.), $p < 0.0001$) supporting the spline model over a linear model.

For birth weight the 5-knot model minimised the AIC but for gestational age there

Table 3.1: Distribution of infant risk factors for death before discharge by survival status

		Survived to discharge		Died		Total
		n	%	n	%	n
Gestational age (weeks)	22	4	16.7	20	83.3	24
	23	147	39.3	227	60.7	374
	24	539	61.3	341	38.8	880
	25	773	77.8	220	22.2	993
	26	1108	83.8	214	16.2	1322
	27	1377	87.5	197	12.5	1574
	28	1932	92.4	158	7.6	2090
	29	2254	95.8	99	4.2	2353
	30	2837	97.2	81	2.8	2918
	31	3754	97.7	87	2.3	3841
Birth weight (g)	< 500	52	39.7	79	60.3	131
	$500 - 749$	1395	67.7	665	32.3	2060
	$750 - 999$	2848	85.8	470	14.2	3318
	$1000 - 1249$	3277	94.1	205	5.9	3482
	$1250 - 1499$	3233	96.3	124	3.7	3357
	$1500 - 1749$	2579	97.8	57	2.2	2636
	$1750 - 1999$	1061	97.4	28	2.6	1089
	$2000 - 2249$	219	95.2	11	4.8	230
	$\geqslant 2250$	61	92.4	5	7.6	66
Birth weight z-score*	< -2	667	82.3	143	17.7	810
	-2 to -1	2390	88.8	300	11.2	2690
	-1 to 0	5807	89.3	699	10.7	6506
	0 to 1	4968	92.6	398	7.4	5366
	1 to 2	786	93.0	59	7.0	845
	> 2	103	80.5	25	19.5	128
	Missing	4	16.7	20	83.3	24
SGA†	No	12551	90.5	1315	9.5	13866
	Yes	2170	87.5	309	12.5	2479
	Missing	4	16.7	20	83.3	24
Sex	Female	6681	91.0	661	9.0	7342
	Male	8044	89.1	983	10.9	9027
Steroids‡	No	1585	81.2	366	18.8	1951
	Yes	12945	91.2	1256	8.8	14201
	Missing	195	89.9	22	10.1	217

* BWT z-score calculated using the RCPCH reference Wright et al. (2010); note this does not extend to infants born below 23 weeks gestation so z-score is missing for these infants

† SGA Small for gestational age

‡ Any dose of antenatal steroids given to the mother

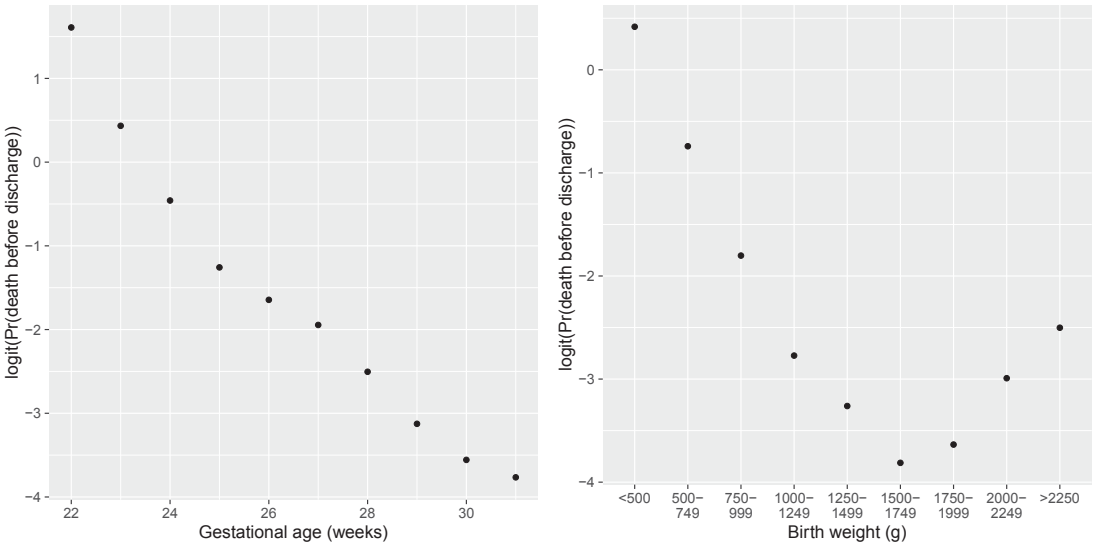


Figure 3.2: Non linearity of gestational age and birth weight with log odds of death

	Akaike Information Criterion	
	Gestational age	Birth weight
3 knots	8525.5	8819.2
4 knots	8525.1	8814.5
5 knots	8524.1	8805.2

Table 3.2: Akaike Information Criteria to compare number of knots

was little difference. Plots of predicted probabilities suggested the 5-knot spline could be replaced with a more parsimonious quadratic model (i.e. with linear and squared birth weight). This gave a lower AIC of 8799.3, but plots of predicted probabilities indicated that the quadratic term may force curvature in the extremes of the data, as indicated in figure 3.3. Both the 5-knot spline and the quadratic model fitted the data better than a linear association alone ($\chi^2 = 278$ on 3 d.f., $p < 0.0001$; $\chi^2 = 297$ on 3 d.f., $p < 0.0001$ respectively). A spline fitted to the quadratic term showed no improvement in fit, therefore there was no evidence of non-linearity beyond the quadratic model. I therefore developed separate models specifying the association with birth weight both ways with a view to comparing the final models.

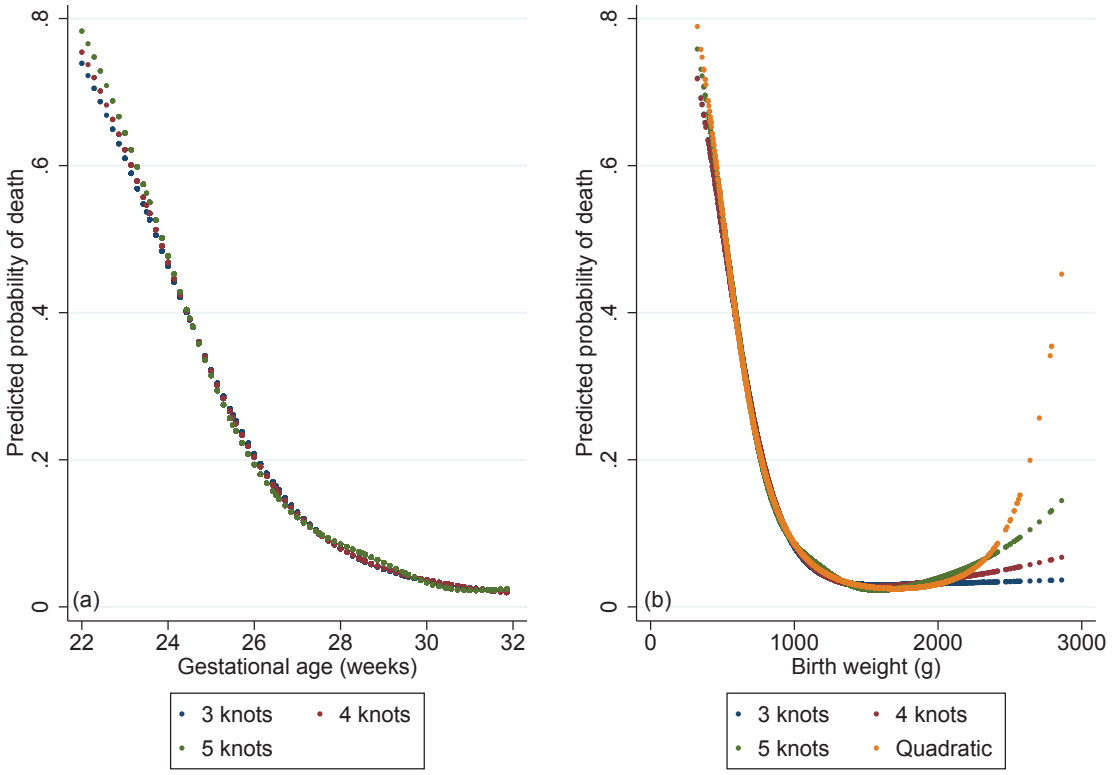


Figure 3.3: Predicted probabilities from logistic regression models for (a) gestational age and (b) birth weight, using 3, 4 and 5 knot splines

Functional form for interactions

First I considered the full model with a 3-knot spline for gestational age, a 5-knot spline for birth weight, and with sex of the infant and whether the mother received any antenatal steroids as binary variables. Interactions between all gestational age and birth weight terms were included. This is the *full* interactions model in table 3.3. Evidence supporting the inclusion of interactions involving only non-linear terms (G_2B_2, G_2B_3, G_2B_4) was weak ($p = 0.073$). These were removed to give the model with

only linear×linear and linear×spline interactions (the *linear*interactions model in table 3.3). From this model, there was no evidence to support the inclusion on any interactions involving non-linear terms ($G_1B_2, G_1B_3, G_1B_4, G_2B_1$) either considered together ($p = 0.43$) considering separately those non-linear in gestation ($p = 0.34$) and birth weight ($p = 0.31$). This resulted in a model with only the linear×linear interaction, for which there was strong evidence for inclusion ($p < 0.001$). The final spline model is given by model (3) in table 3.3.

A similar procedure was carried out modelling birth weight with linear and quadratic terms; results from regression models are shown in table 3.4. Again there was little evidence from the full model (table 3.4 *full* interactions model) to keep the interaction with only non-linear terms ($G_2B^2, p = 0.11$) resulting in the *linear* interactions model in table 3.4 . Removal of further interactions did not improve the fit of the model.

The AIC of the final model with birth weight as a 5-knot spline was 8238 compared with 8227 modelling birth weight as quadratic, with 9 covariates in each model. One concern evident in figure 3.3 was that the quadratic shape was forcing the shape of the model and potentially producing poor fit for observations in the tails of the birth weight distribution. I investigated this using residual and leverage plots. Plots of standardised residuals against birth weight and predicted probabilities (figure 3.4) showed no evidence of non-linearity as the smoothed values for the mean (given by locally weighted regression) gave a straight line along zero. Note that unlike linear regression we do not expect a random scatter as all observed values are either zero or one. There was no evidence of a difference between models for particular birth weight or probability ranges based on patterns of residuals (figure 3.4). Some observations had disproportionate leverage in the quadratic model (figure 3.5) compared with the spline model (figure 3.6). The predicted probability of death was around 0.2 for these observations and they were for infants of high birth weight and gestation approaching 32 weeks. The residuals were small, indicating that they survived (as the observed outcome was close to the predicted probability) and that the model fits reasonably well for these observations. A sensitivity analysis was run excluding observations with leverage over 0.04, which gave almost identical results to the model on the whole dataset. I therefore selected the quadratic model as the final model for the data as it gave the lowest AIC. The odds ratios for the final model and the sensitvity analysis are shown in table 3.5.

Figure 3.7 illustrates survival probabilities relative to gestational age and birth weight. Probabilities depend on whether the infant is male or female and whether the mother was given antenatal steroids, therefore the resulting charts will differ. Figure 3.7 shows two charts: 3.7(a) for females whose mothers were given antenatal steroids and 3.7(b) for males whose mothers were not given antenatal steroids, with the former having better survival propsects as indicated by the odds ratios in table 3.5. Survival

	Full	Linear	Simple
Male	0.35	0.34	0.34
	(0.23,0.46)	(0.22,0.46)	(0.22,0.45)
Steroids	-0.83	-0.83	-0.83
	(-0.98,-0.68)	(-0.98,-0.68)	(-0.98,-0.68)
G_1	-1.21	-1.69	-1.83
	(-1.96,-0.46)	(-2.21,-1.18)	(-2.18,-1.49)
G_2	-0.52	0.96	0.78
	(-2.17,1.13)	(0.45,1.48)	(0.47,1.10)
B_1	-1.87	-2.40	-2.44
	(-2.71,-1.02)	(-2.98,-1.83)	(-2.83,-2.04)
B_2	4.74	8.30	9.08
	(-1.38,10.9)	(4.75,11.8)	(6.74,11.4)
B_3	-10.9	-23.2	-24.5
	(-33.5,11.8)	(-35.7,-10.8)	(-32.6,-16.4)
B_4	10.1	29.3	26.7
	(-24.7,44.9)	(11.6,46.9)	(15.5,37.9)
G_1B_1	-0.24	-0.58	-0.66
	(-0.76,0.29)	(-0.94,-0.22)	(-0.86,-0.46)
G_1B_2	-4.04	-1.22	
	(-9.40,1.32)	(-4.21,1.78)	
G_1B_3	17.9	3.24	
	(-10.1,45.8)	(-6.85,13.3)	
G_1B_4	-49.4	-5.39	
	(-117.9,19.2)	(-21.3,10.5)	
G_2B_1	-0.91	0.24	
	(-2.18,0.35)	(-0.25,0.73)	
G_2B_2	7.15		
	(-0.95,15.3)		
G_2B_3	-25.0		
	(-53.9,3.93)		
G_2B_4	42.3		
	(-6.90,91.4)		
Intercept	-4.23	-4.95	-5.04
	(-5.38,-3.09)	(-5.72,-4.18)	(-5.64,-4.43)

Table 3.3: Comparison of regression coefficients (95% confidence intervals) for models with different gestational age-birth weight interactions; birth weight modelled as a 5-knot spline

	Full	Linear
Male	0.34	0.34
	(0.23,0.46)	(0.23,0.46)
Steroids	-0.83	-0.84
	(-0.98,-0.68)	(-0.99,-0.69)
G_1	-2.26	-2.11
	(-2.79,-1.74)	(-2.60,-1.61)
G_2	1.18	1.09
	(0.73,1.63)	(0.66,1.53)
B	-0.36	-0.23
	(-0.63,-0.080)	(-0.45,-0.0038)
B^2	0.57	0.67
	(0.41,0.74)	(0.55,0.78)
G_1B	-1.61	-1.26
	(-2.27,-0.94)	(-1.77,-0.75)
G_1B^2	-0.43	-0.26
	(-0.68,-0.19)	(-0.38,-0.14)
G_2B	0.57	0.35
	(0.12,1.02)	(0.015,0.69)
B^2G_2	0.14	
	(-0.031,0.31)	
Intercept	-3.26	-3.23
	(-3.62,-2.90)	(-3.59,-2.87)

Table 3.4: Comparison of regression coefficients (95% confidence intervals) for models with different gestational age-birth weight interactions; birth weight modelled as quadratic

increases with gestation, with steeper absolute increases at earlier gestations indicated by the closer contours. Survival in general increases with birth weight, though adverse effects of being very high birth weight for gestation can be seen at later gestational ages.

	Final	Sensitivity*
Male	1.41	1.41
	(1.25,1.59)	(1.26,1.59)
Steroids	0.43	0.43
	(0.37,0.50)	(0.37,0.50)
G_1	0.12	0.12
	(0.074,0.20)	(0.074,0.20)
G_2	2.98	2.96
	(1.93,4.61)	(1.90,4.61)
B	0.80	0.80
	(0.64,1.00)	(0.64,1.00)
$B2$	1.95	1.95
	(1.74,2.18)	(1.74,2.18)
G_1B	0.28	0.29
	(0.17,0.47)	(0.17,0.49)
G_1B^2	0.77	0.77
	(0.68,0.87)	(0.68,0.88)
G_2B	1.42	1.41
	(0.99,2.04)	(1.02,1.99)

* Excluding six points with high leverage ($\geqslant$0.04)

Table 3.5: Comparison of odds ratios (95% confidence intervals) for final model and sensitivity analysis excluding influential observations

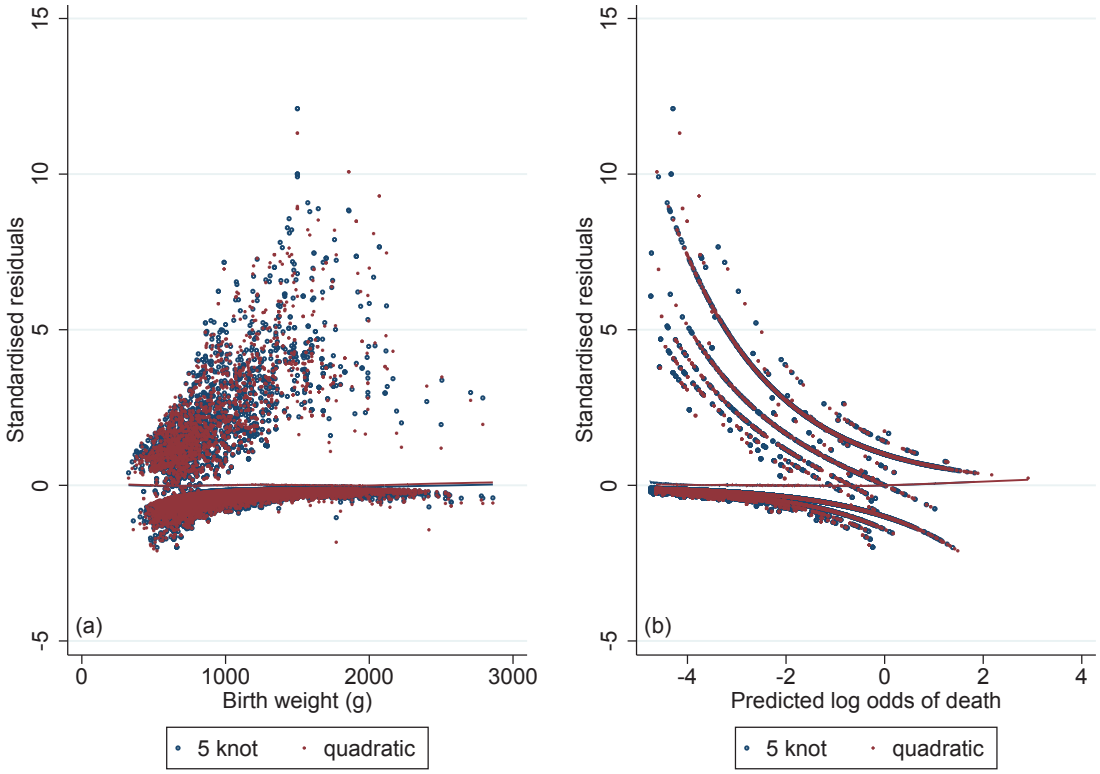


Figure 3.4: Standardised residuals against (a) birth weight and (b) predicted probabilities comparing spline and quadratic models for birth weight

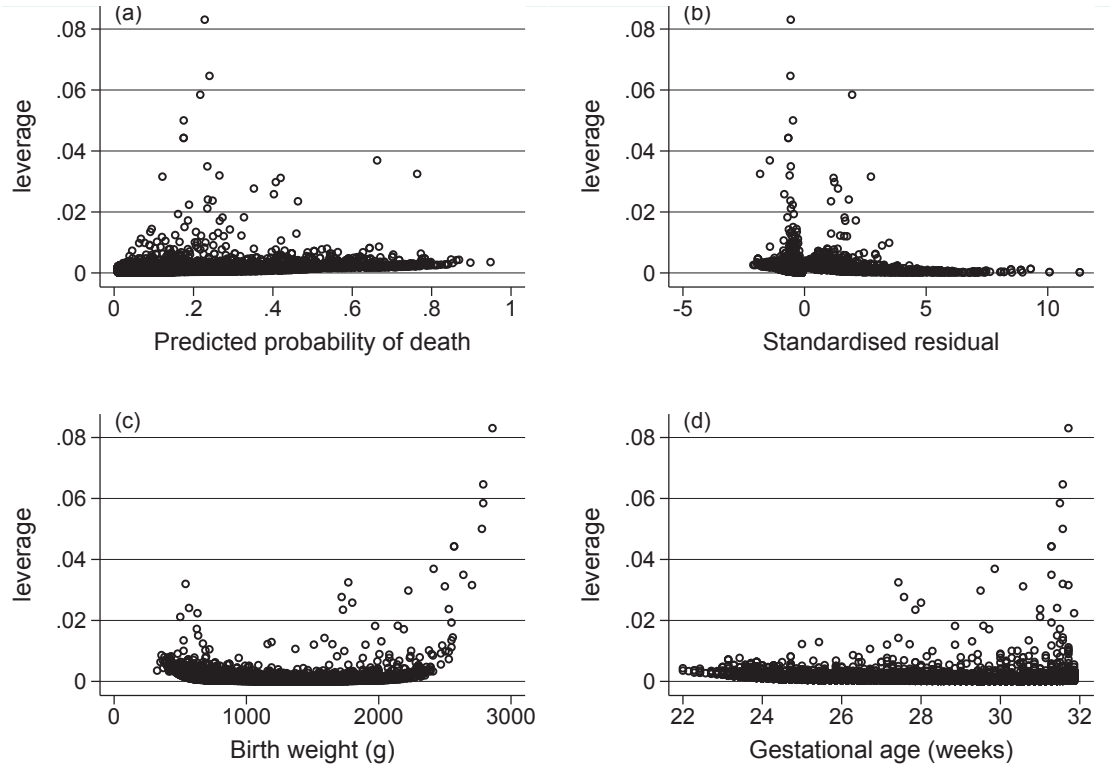


Figure 3.5: Leverage from quadratic model against (a) predicted probabilities (b) standardised residuals (c) birth weight and (d) gestational age

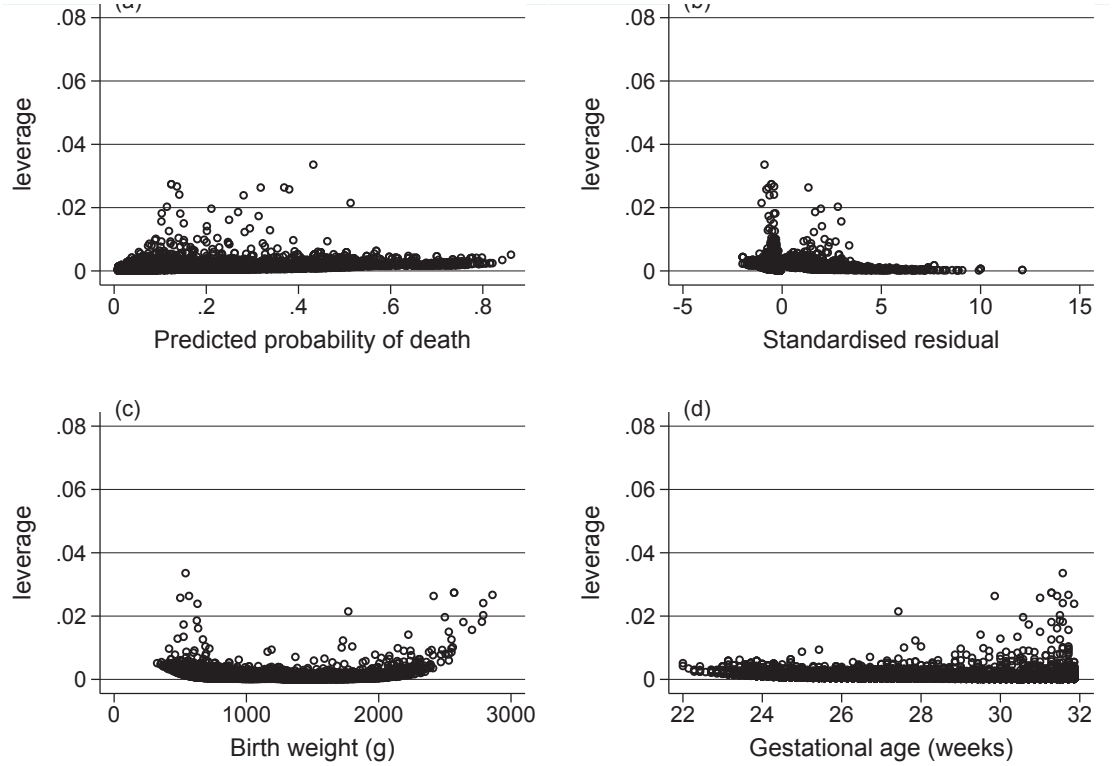


Figure 3.6: Leverage from spline model against (a) predicted probabilities (b) standardised residuals (c) birth weight and (d) gestational age

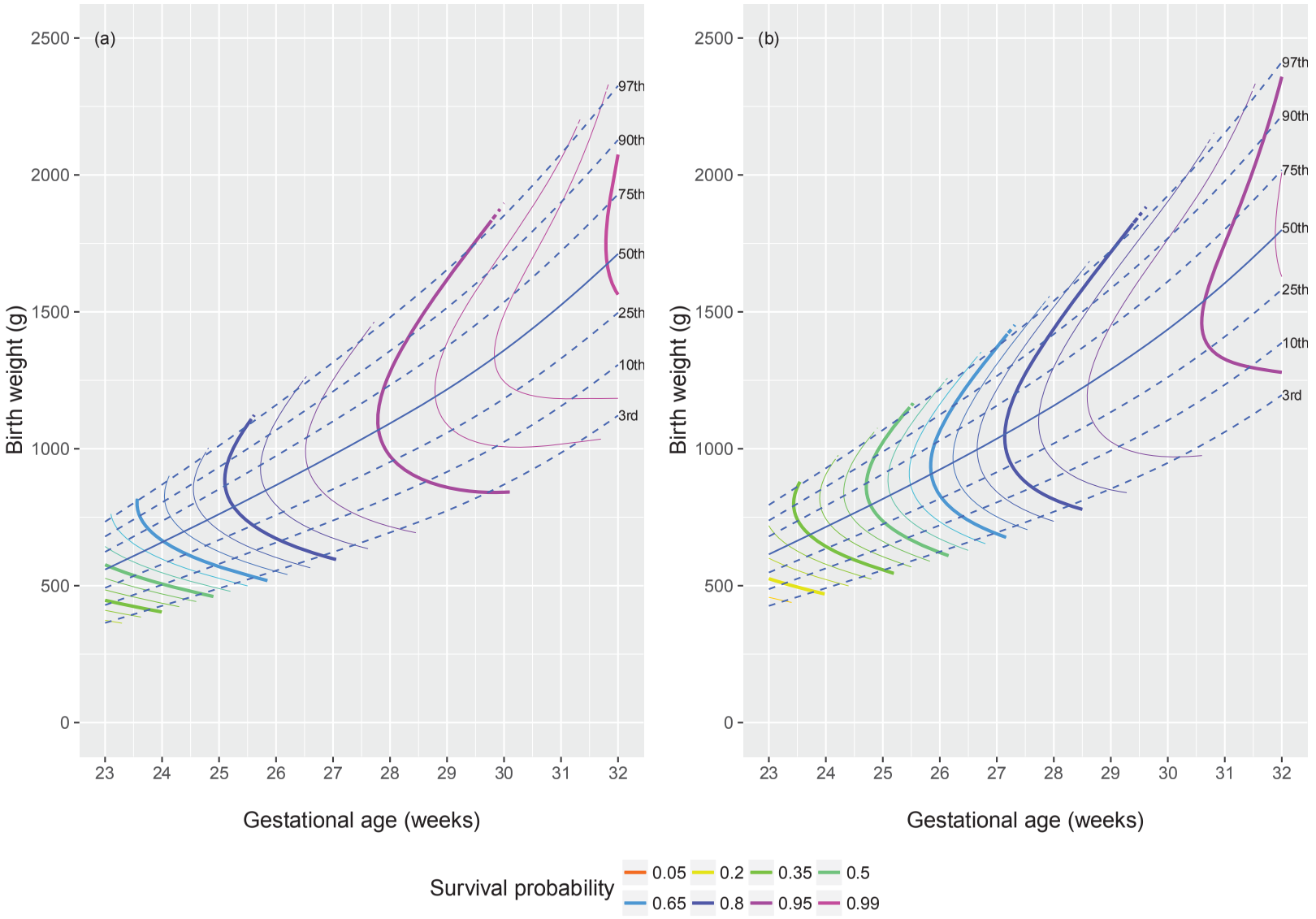


Figure 3.7: Survival probabilities for singleton infants by gestational age and birth weight for (a) females whose mothers were given antenatal steroids (b) males whose mothers were not given antenatal steroids. Birth weight centiles are shown as dotted lines. Minor contours are equally spaced between the major contours shown in the legend.

Model performance Table 3.6 shows the model performance statistics before and after correcting for optimism. The model showed good predictive performance (c-statistic above 0.8, Brier scores and calibration intercepts close to zero and calibration slopes close to one) with little difference between the apparent and optimism-corrected estimates.

	Apparent	Corrected
c-statistic	0.8214	0.8205
Brier score	0.0725	0.0727
Infants $< 28^{+0}$ weeks gestation		
Calibration slope	1.028	1.030
Calibration intercept	0.023	0.025
Infants $\geqslant 28^{+0}$ weeks gestation		
Calibration slope	0.971	0.957
Calibration intercept	-0.077	-0.117

Table 3.6: Comparison of apparent and optimism-corrected performance statistics

Model performance statistics compared with three other models are shown in table 3.7. For all models the predicted mortality for the NNRD model was closed to the observed mortality than that predicted by the other models. Discrimination (indicated by the c-statistics) and overall fit (Brier score) were generally fairly similar comparing the NNRD model with both he TNS and CRIB II models, with slightly better results compared with the NICHD model. The NNRD was in general better calibrated than the other models, though predictions for the NICHD model were under-dispersed. The comparator models all gave predictions that were too high and over-dispersed.

Table 3.7: Comparison of NNRD model performance with: (i) the Trent Neonatal Survey model (TNS) (ii) National Institutes of Child Health and Development model (NICHD) and (iii) the updated Clinical Risk Index for Babies score (CRIB II)

	TNS		NICHD		CRIB II	
	TNS	NNRD	NICHD	NNRD	CRIB II	NNRD
N	10900		2252		6002	
Obs. mortality n(%)	1055 (9.7)		799 (35.5)		529 (8.8)	
Pred. mortality (%)	10.6	9.9	46.2	36.3	7.7	8.5
c-statistic	0.81	0.82	0.60	0.72	0.81	0.80
	(0.80,0.83)	(0.81,0.83)	(0.58,0.63)	(0.69,0.74)	(0.80,0.83)	(0.78,0.82)
Brier score	0.072	0.072	0.242	0.198	0.068	0.067
Calibration slope	0.88	1.00	0.48	1.24	0.76	1.00
	(0.83,0.93)	(0.95,1.04)	(0.36,0.59)	(1.09,1.39)	(0.70,0.83)	(0.92,1.08)
Calibration intercept	-0.32	0.00	-0.54	0.09	-0.26	0.05
	(-0.42,-0.22)	(-0.10,0.09)	(-0.62,-0.45)	(-0.04,0.21)	(-0.42,-0.09)	(-0.13,0.23)

3.5 Discussion

3.5.1 Summary

In this chapter I selected the covariates for risk adjustment. Transfers are a crucial aspect of the thesis and known to be associated with mortality in a number of ways, so were also considered as a risk factor but rejected as transfer management is an integral part of care provided by both NNUs and networks. I considered level of NNU and volume of care as possible provider-level characteristics. I selected only level of NNU as this is an objective, pre-assigned measure whereas volume may be influenced by care through increased referrals. Sociodemographic data such as ethnicity were not considered for inclusion due to data quality, and because doing so can reduce the ability to draw desired inferences about providers. Established clinical risk factors used in existing models were gestational age, birth weight, average size for gestation, sex, ethnicity, antenatal steroids, Apgar score, temperature on admission, measures of respiratory status. Of these, Apgar score, temperature on admission, measures of respiratory status were not selected because they may be influenced by initial neonatal care provided on the labour ward, rendering these variables inappropriate for provider profiling purposes. Average size for gestation is a reflection of birth weight for gestation, so I chose to model this as an interaction so as not to lose information by categorisation. I then described these variables in the NNRD data and their associations with mortality. I modelled the interacting influences of gestational age and birth weight on mortality using restricted cubic splines. The model showed good discrimination and calibration compared with other existing prediction models.

3.5.2 Selection of risk factors

I chose to select variables based on the review by Medlock et al. (2011) rather than conduct a repeat review of the literature. This study also formed the basis for risk adjustment for other work produced by the Neonatal Data Analysis Unit (Modi et al., 2016; Watson et al., 2014). The search strategy of Medlock et al. seemed adequate, the review included all studies identified by other earlier (non-systematic) reviews, and no further variables were found by updating the search. Therefore, while risk adjustment models are never perfect it seems unlikely that any major clinical predictors have been missed for the purposes of this thesis. Nonetheless there is the potential for inconsistency in the variables selected. I only considered the variables deemed *frequent predictors* by Medlock et al.. The authors did not give their exact criteria for this regarding inclusion in final prediction models, though they did state that those variables had to be tested in at least ten studies. There were several other predictors which were examined in

several studies but not classed as frequent predictors including maternal health risks such as pregnancy-related hypertensive disorders. Application of different criteria could have resulted in different variables being selected. Furthermore individual studies used different methods for variable selection, and smaller studies might not have sufficient power to detect small effects. An alternative and perhaps more comprehensive approach might be to include all candidate variables in the risk adjustment model provided they meet the criteria described in section 3.2

The factors we adjust for affect the interpretation of the results and to what purposes they can be used. The term "provider profiling" covers analyses with a wide range of purposes, but broadly refers to the estimation, assessment or comparison of performance indicators across multiple providers. In this thesis I am primarily concerned with estimating risk-adjusted mortality rates for individual NNUs, rather than testing for outliers or comparing NNUs with each other (the rationale for this approach is driven by the data available and the methods used, and will be discussed in chapter 4). As set out in the aims of the thesis (section 1.4) the purpose of estimating these rates is to inform NNUs how the mortality for infants in their care compares with what we would expect to see for similar infants attending similar NNUs. The choice of risk factors governs the interpretation of "similar". In this thesis I have adjusted for established clinical risk factors but not sociodemographic variables for reasons discussed in sections 3.1 and 3.2, so the results are not adjusted for all the risk factors which are not under the control of the provider. This means that if mortality at a particular NNU appears high or low compared with what we expect, this is not necessarily caused by the provider as we cannot rule out the influence of minor clinical risk factors or the sociodemographic profile of the patient population. The analysis is therefore useful as a starting point for further investigation, rather than a conclusive estimate of performance. The current situation is that NNUs have no provider-specific mortality data which includes transferred patients, so another use of the analysis is simply to provide information which was not previously available. In chapter 8 I show how the results of the thesis could be presented in a case study for a single NNU and describe how results could be used by clinical and managerial staff.

3.5.3 Statistical methods

There were 217 infants excluded because of missing data for antenatal steroids. As discussed in section 2.2.7, 341 infants were already excluded because of missing outcome data. The number of these infants is small and those with missing steroids data had a mortality rate identical to the overall cohort so the impact of the missing data is likely to be small. Techniques for handling missing data have been applied in some provider profiling studies (Glance et al., 2009; Tekkis et al., 2003) and I will discuss how these

can be incorporated into methods developed in this thesis in chapter 9.

I used restricted cubic splines to model the non-linear associations. Gestational age can be considered as a categorical variable by week (or groups of weeks). While this gives clear interpretation to the resulting parameters, given that a single week can make such a big difference to mortality, rounding up or down by a few days is non-negligible, and would be an unnecessary loss of information since the data are available. Use of higher order polynomials (e.g. $logit(p_i) = \alpha + \beta_1 x_i + \beta_2 x_i^2$) is a common option for non-linear associations and relatively easy to interpret (Cole et al., 2010), with a more parsimonious if less interpretable variant in the form of fractional polynomials (Manktelow et al., 2013; Royston and Sauerbrei, 2008) which allow non-integer powers. These methods both have the disadvantage that the shape in some areas (possibly only determined by a few data points) can force an incorrect relationship in other areas of the distribution because the same functional form is assumed, as was investigated here for the quadratic birth weight model. Another possible approach generalised additive models, which describe complex relationships using sums of smooth functions of the predictor of interest (Hastie and Tibshirani, 1986). The functions can be non-parametric, so these models also offer flexibility without forcing the shape elsewhere. Initial exploration using generalised additive models produced similar predictions but a more thorough investigation is required to establish whether this method presents a superior alternative.

3.5.4 NNRD data and associations with mortality

The validation exercise comparing NNRD data with that recorded independently for a clinical trial (see section 2.1.3) found that data quality for the risk factors used here was generally good with regard to differences that were classed as *major discrepancies*. The proportion of infants with a major discrepancy recorded was 3% for gestational age (one week or more), 0.9% for birth weight (100g or more), 0.2% for sex and 2.4% for antenatal steroids. However 20.4% of infants had discrepancies of three to six days in their recorded gestational age (up to two days was considered acceptable). In addition three of the 1258 infants had a different survival outcome recorded. Errors can lead to biased results in provider profiling studies if they are concentrated in particular providers. I have not taken any further action in this thesis regarding checking for errors, but as described in chapter 2, NNUs are given the opportunity to check and correct the data which feeds into regular audit reports.

Gestation-specific mortality rates are very similar to those for neonatal mortality for livebirths described in section 3.3.2. While this is not evidence of completeness of the population, it is reassuring with regard to the potential for bias. Mortality was slightly

lower at earliest gestations compared with Office of National Statistics data, possibly because deaths soon after labour (prior to admission to neonatal care) are more likely to occur at this age (see section 2.2.6). As described in section 1.1.3, mortality rates for preterm infants, particularly those at the limits of viability, can vary across and even within countries due to differences in resuscitation practices and other clinical management decisions. The most similar cohort to that presented here come from the EPICure 2 study (Costeloe et al., 2012) which found that of all infants born in England in 2006 between 22^{+0} and 25^{+6} weeks gestation, 47% died before discharge. This figure is considerably higher than the 36% found here, reflecting progress in survival of extremely preterm infants over the past decade. Patterns of association between birth weight and mortality are also similar to those described in the literature for all births. Mortality decreased with birth weight, with the lowest mortality for infants with a z-score of between 1 and 2, followed by an increase in mortality for those at the highest birth weights.

3.5.5 Model performance and validation

There were minimal differences before and after optimism correction, possibly because the sample size was large. Bootstrapping is advantageous over split-sample validation because all the data can be used for developing the model. Although the overall sample was large, there are few infants born at lower gestations and few deaths at the upper gestations so using all available data was preferable. As stated in section 2.2.1 I did not use data outside 2011-2013 because of incomplete population data prior to this period and changes to the structure of neonatal services afterwards (these changes will be discussed further in 9).

A strength of the analysis was that variable selection was based on evidence from the literature rather than model-based selection, which can result in overfitting. Functional form and interactions for gestational age and birth weight were determined as part of the modelling process. This procedure could have been made more robust by applying bootstrapping techniques to model selection as well as to correct for over-optimisim. This is feasible for automated model selection procedures but more difficult otherwise, particularly if graphical assessment is used. As the NNRD is constructed from an ongoing data collection a better way to validate all aspects of the model would check its predictive properties on an independent dataset from a later time period.

The probabilities of death cover a wide range and gestational age is very discriminatory, which means that the *c*-statistic will be high for any model incorporating gestational age. This is supported by the *c*-statistic being lower for the models which use smaller subgroups. The model performed well against comparator prediction models.

The discriminatory power of existing models was good but calibration was poor and the models in general over-predicted mortality. This could be because of population differences but also because survival of preterm infants has improved since they were developed. Comparator models also all gave over-dispersed values which is perhaps more surprising as this result is usually an indication of over-fitting (Steyerberg et al., 2010), but these models were not fitted to this dataset. A possible explanation for this is that as the reductions in mortality seen over time have by and large been for the lowest gestations (section 1.1.3), so the variability in predicted mortality will be reduced, resulting in less dispersed predictions. The comparator models might have performed better if re-calibrated to this population (Manktelow et al., 2010), presenting an alternative approach to developing a model internally.

The results of this chapter provide the risk adjustment strategy to be adopted for the rest of the thesis. In the next chapter I review methods for provider profiling; the risk-adjustment developed here will be incorporated into the methods arising from this review in chapter 5.

Chapter 4

Review of hierarchical models and provider profiling methods

The aim of this chapter is to summarise existing methods relevant to the research questions of this thesis. In chapter 1, I introduced provider profiling and briefly described how a class of statistical models known as hierarchical models are often used for such studies. In section 4.1, I define hierarchical models and consider their use in provider profiling studies generally and for this thesis. This is followed in section 4.2 by a systematic review of the literature to establish how patient transfers are handled in existing provider profiling studies. I then introduce multiple membership models as a potential method for allowing for transfers, and review their applications in the literature (section 4.3). Finally, in section 4.4 I will discuss alternative approaches and relate the findings of the chapter to the aims of the thesis.

4.1 Hierarchical models

The aim of this section is to define a basic logistic hierarchical model and some of its extensions, review the use of hierarchical models for provider profiling, and then discuss the extent to which they can be used to answer the research questions of the thesis.

4.1.1 Definition

Observations often fall into groups, as seen in both in analysis of healthcare data (patients at the same GP surgery) and elsewhere (pupils in the same class, employees at the same company, piglets in a litter). Observations in the same cluster are likely to

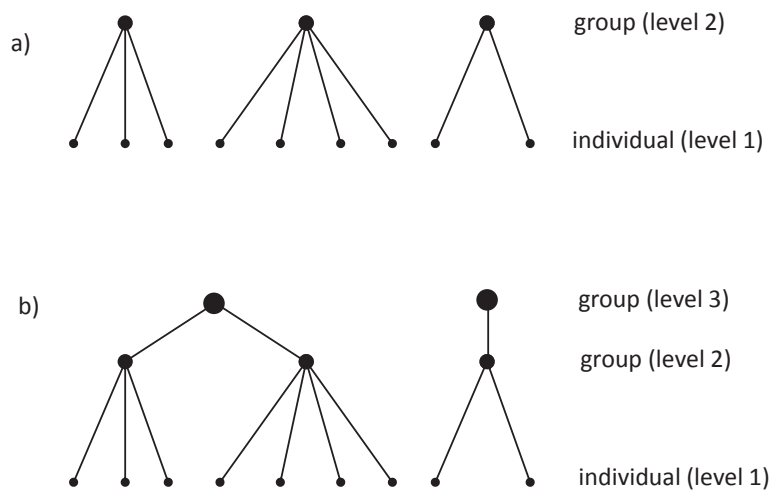


Figure 4.1: Diagrams illustrating hierarchical data structures: (a) a two-level structure of individuals within groups (b) As figure 4.1(a) but with the three levels, where each group at level 2 is a member of a group at level 3

be correlated (e.g. due to peer group influence in a class), but standard statistical models assume that all the observations are independent. Ignoring the structure of the data by assuming independence leads to standard errors which are underestimated for cluster-level variables and overestimated for individual-level variables (Rabe-Hesketh and Skrondal, 2005). Hierarchical models (also called *random effects* or *multilevel* models) provide a way of correctly accounting for dependencies in the data. It is said that the observations are *nested* in *clusters* or *levels*. Figure 4.1(a) illustrates a basic two-level structure. What follows is a brief introduction to hierarchical models; further details can be found in standard texts such as Goldstein (2002); Rabe-Hesketh and Skrondal (2005); Snijders and Bosker (1999).

Suppose we are interested in an outcome Y measured in N individuals and we wish to adjust for some continuous covariate X. For consistency with the research question for the thesis let Y be binary, and denote the probability of an individual i experiencing the outcome as p_i. As described when developing the risk adjustment model in section 3.4, we can model the probability as a function of the covariate using logistic regression:

$$logit(p_i) = log(\frac{p_i}{1-p_i}) = \alpha + \beta x_i \tag{4.1}$$

In equation 4.1, α is the log odds of Y when $X = 0$ and β is the increase in the log odds of Y corresponding to a one unit increase in X. We can obtain the odds ratio for the effect of X by exponentiating β. Suppose these N individuals are nested within J clusters, for example patients treated at different clinics. If we expect two patients in the

same clinic to have more similar values of Y than two patients in different clinics, then our observations are not independent. We can model the effect of the cluster with an additional term u_j and the assumption that the u_j arise from a common distribution, for example Normal with mean 0 and variance σ_u^2. We also assume that the u_j are uncorrelated with covariates x_{ij}.

$$\begin{aligned} logit(p_{ij}) &= \alpha + \beta x_{ij} + u_j \\ u_j &\sim N(0, \sigma_u^2) \end{aligned} \tag{4.2}$$

We use the subscript ij to denote individual i in cluster j. The u_j is the increase in the log odds of Y associated with cluster j. The u_j can then be exponentiated to obtain an odds ratio. This odds ratio is interpreted as the odds of Y for an individual in cluster j compared with the odds of Y for an individual (with the same value of X) in an average cluster. There is a slight change in the interpretation of the other parameters: α is the log odds of Y when $X = 0$ and $u_j = 0$, and β is the increase in the log odds of Y corresponding to a one unit increase in X for individuals in the same cluster. The cluster-specific interpretation of β in equation 4.2 is not the same as the covariate effect in the population as a whole (the interpretation in equation 4.1). This can be obtained using a *population-averaged* (PA) model (also called a *marginal* model) (Szmaragd et al., 2013). Population-averaged models present an alternative to hierarchical models for analysing clustered data in some circumstances. In PA models the similarity between individuals in the same cluster is not described using model parameters (the u_j and σ_u^2 in equation 4.2), but is specified using a working correlation structure for the residuals which describes the pattern of correlation with clusters. The correlation is considered a nuisance parameter, rather than an aspect of the model that is of specific interest. The between-cluster variance and cluster effects cannot be estimated, so these models are not suitable for the thesis and are not considered further here.

This model is often called a *random intercept* model as it is equivalent to allowing the intercept term to vary randomly across clusters; in model 4.2 the intercept is decomposed as a fixed part α and a random part u_j, with the u_j called *random effects*. Other continuous or categorical covariates can be added to adjust for other factors either at the individual level or the cluster level. The assumption of a common distribution reflects the fact that the clusters are in some sense similar: if asked to guess the post-operative morality rate for a surgeon, we might base the guess on the rates for other surgeons, and perhaps close to other surgeons in the same hospital. It also means that we only need to estimate one additional model parameter, σ_u^2, which tells us about the amount of variation at cluster level. Without this assumption we would have to estimate each u_j independently, known as a *fixed effects* model. This may be infeasible if there many clusters relative to the number of observations, and if some clusters are small the estimates will be poor. In addition, fixed effects model cannot be used to

model more than two levels of the hierarchy, accommodate cluster-level covariates, or model variation in covariate effects across clusters (see section 4.1.3 for how these can be achieved using hierarchical models). However fixed effects models do not require the assumption that the u_j are uncorrelated with the other covariates.

In practice normal random effects are the default option, though this is a reasonable choice when the residual variation between clusters is the result of small variations, as the sum of these will be approximately normal (Ohlssen et al., 2007b). The t distribution is similar to a Normal but with heavier tails and can be used in place of a Normal distribution in the presence of outliers. In addition to the mean and standard deviation, the distribution is determined by an additional parameter (the degrees of freedom, here denoted ν). As ν increases, the t distribution tends towards Normal. The t distribution accommodates outliers when the distribution is symmetric, but there may be situations where there are outliers on one side of the distribution, which could be modelled by a skewed distribution (Lee and Thompson, 2008). Other less restrictive distributions are also possible, such as mixtures of Normals (Austin, 2009) or flexible distribution using non-parametric techniques (Ohlssen et al., 2007a).

4.1.2 Model estimation

Hierarhical models can be estimated in a number of different ways. Whether we wish to use a *frequentist* or *Bayesian* approach is one consideration affecting estimation method. Frequentist (also called *classical*) methods are often used as a default as they provide analytical solutions for most standard estimation problems which is not the case for Bayesian methods. The logistic regression models in chapter 3 were estimated in this way. More detail on Bayesian methods including basic theory will be provided in section 5.1, but the fundamental difference is how we view probability. The Bayesian view of probability is a degree of certainty, so we can assign a probability distribution to anything including quantities for which an unknown true value exists (e.g. mean birth weight in a population). We can do so regardless of whether we have any data (e.g. a sample of 1000 birth weights). This is not the case under a frequentist view: the mean birth weight is a fixed value and any uncertainty we have is only expressed about the experimental data we have (e.g. the probability that a particular range constructed from our sample of 1000 birth weights contains the unknown true mean).

Taking a frequentist approach, we can fit the model in a variety of ways (e.g. restricted maximum likelihood and iterative generalised least squares). Likelihood methods for binary outcomes entail the use of approximations such as Gaussian quadrature and Laplace approximations. The different methods vary in efficiency, can produce slightly different results and enable different hypothesis tests to be used; further detail

can be found in Rabe-Hesketh and Skrondal (2005). A second stage of estimation is required to obtain the random effects denoted by u_j. The maximum likelihood estimates are obtained by plugging in the estimated model parameters (taking them to be the certain true values) then estimating the cluster effects as unknowns (Rabe-Hesketh and Skrondal, 2005). This is carried out separately for each cluster and therefore cannot be estimated for clusters with no events or with only events (e.g. a hospital where no patients die, or a hospital where all patients die). The alternative approach predicts the cluster effects based on the most likely values for the model parameters, including those of the common distribution of cluster effects. Known as *empirical Bayes*, this method uses Bayesian principles in that the cluster effects are assingned a probability distribution. Unlike full Bayes we are using the data twice (once for the likelihood and once to estimates the parameters of the random effect distribution) though there are ways to correct for this overconfidence (Carlin and Louis, 2008).

We can use full Bayesian methods to fit the model and estimate the random the random effects in one step (Gelman et al., 2004). One difference between frequentist and Bayesian methods which is of relevance here is that all estimated quantities are considered random, unlike frequentist statistics where a distinction is drawn between fixed and random parameters in the model. Hierarchical models therefore fit naturally into a Bayesian framework: the random effects distribution forms part of our prior information (our beliefs before seeing the data) and the posterior distribution (our beliefs after seeing the data) for the random effects are estimated in a similar manner to other model parameters. Further explanation of prior and posterior belief, and the mechanics of estimation, will be discussed in section 5.1. Whether empirical or full Bayes is used, the resulting cluster effects are shrunk towards the mean of their distribution (in this case zero) in contrast to the maximum likelihood estimates. The extent of the shrinkage depends on how much data we have relative to the prior, with smaller clusters experiencing more shrinkage. This is intuitive for some purposes: if estimating lung cancer rates by area, in sparsely populated areas we might be more inclined to look to data from neighbouring areas than in densely populated ones.

4.1.3 Extensions to the basic model

Individuals can be similar at more than one level (e.g. surgery patient i within surgeon j within hospital k), with similarity now also occurring at the middle level(s) (e.g. similarity of surgeons within the same hospital in the previous example). This structure, illustrated in figure 4.1(b), can be modelled by adding another random effect to the

model specification:

$$\begin{aligned} logit(p_{ijk}) &= \alpha + \beta x_{ijk} + u_{jk} + v_k \\ u_{jk} &\sim N(0, \sigma_u^2) \\ v_k &\sim N(0, \sigma_v^2) \end{aligned}$$

Note that we are now using the index ijk to accommodate the third level. As assumption of independence from covariates x_{ij} now applies to both the u_{jk} and the v_k, and further we assume that the u_{jk} and the v_k are uncorrelated. We can reparametrise the model to consider each level separately:

$$\begin{aligned} logit(p_{ijk}) &= \beta x_{ijk} + \mu_{jk} && \text{(Level 1)} \\ \mu_{jk} &= \nu_k + u_{jk} && \text{(Level 2)} \\ \nu_k &= \alpha + v_k && \text{(Level 3)} \\ u_{jk} &\sim N(0, \sigma_u^2) \\ v_k &\sim N(0, \sigma_v^2) \end{aligned} \tag{4.3}$$

This parametrisation, known as *hierarchical centring*, emphasises the hierarchical structure of the data. There is a shift of focus, as rather than having one equation for the individual we have a set of equations, one for each level of the hierarchy. This approach is intuitive if we wish to add cluster-level variables into the model. If we have some factor Z measured at level 2 of the hierarchy (say an area-level measure of air pollution if examining lung cancer rates) then the estimated log odds of the outcome at hospital j is given by $\mu_{jk} = \nu_k + \gamma z_{jk} + u_{jk}$, i.e. it is a function of the cluster-level covariate.

In model 4.2 all cluster effects arise from the same distribution with a single variance parameter σ_u^2. It is possible to have different variances (known as *complex variation*, see Goldstein (2002)) if we expect the influence of the cluster on the outcome to be heterogeneous. For example, in an educational context, small and large classes could be assigned a different variance. Another common extension is to allow covariate effects (β in equation 4.2) to vary with cluster as well (for example we might expect the effects of low income on lung cancer rates to vary by area). This is illustrated in equation 4.4 below.

$$\begin{aligned} logit(p_ij) &= \alpha + \beta x_{ij} + u_j^{(1)} x_{ij} + u_j^{(0)} \\ \boldsymbol{u} = \begin{pmatrix} u^{(0)} \\ u^{(1)} \end{pmatrix} &\sim N(0, \boldsymbol{\Psi}) \quad \boldsymbol{\Psi} = \begin{pmatrix} \sigma^2_{u^{(0)}} & \sigma_{u^{(0)}u^{(1)}} \\ \sigma_{u^{(1)}u^{(0)}} & \sigma^2_{u^{(1)}} \end{pmatrix} \end{aligned} \tag{4.4}$$

Here, $\boldsymbol{\Psi}$ is the 2 x 2 variance covariance matrix for the random effects. The diagonal terms of the matrix contain the variance of the random intercepts and coeffi-

cient ($\sigma^2_{u^{(0)}}$ and $\sigma^2_{u^{(1)}}$ respectively), while the off-diagonal terms give the covariance $\sigma_{u^{(0)}u^{(1)}} = \sigma_{u^{(1)}u^{(0)}}$ which describes how the random terms are correlated. For example, areas with higher than expected rates of lung cancer may also have stronger associations between income and lung cancer, leading to a stronger correlation between intercept and slope terms. These models are sometimes referred to as *random coefficient models*.

4.1.4 Hierarchical models for provider profiling

The hierarchical models described in the previous section are widely used in public health (Diez-Roux, 2000; Leyland and Goldstein, 2001) and have long been advocated for provider profiling purposes (DeLong et al., 1997; Goldstein and Spiegelhalter, 1996; Normand et al., 1997). In this context, the random effects described above are called *provider effects*, terminology which I shall use in this thesis. These cannot necessarily be interpreted as an effect that is wholly caused by the provider (such that would be obtained if patients were randomised to providers) because the interpretation depends on the set of factors we have adjusted for. This set may not encompass all variables beyond provider control, thus some of the between-provider variation may be due to these unmodelled variables as discussed in section 3.5.2.

Early applications include Gatsonis et al. (1993, 1995) who described variation in angiography rates across states, and Thomas et al. (1994) who estimated hospital-specific log odds of mortality. Gatsonis et al. (1995) also allowed the provider effects to vary with patient characteristics using the random coefficient formulation previously described. These authors recommended the hierarchical analysis over the fixed effects approach because it produces more reliable and precise provider-specific estimates (particularly for smaller providers) through shrinkage, addresses problems of multiple comparisons, accounts for uncertainty in a statistically principled way, and enables attribution of variation to different sources, including provider-level variables (DeLong et al., 1997; Gatsonis et al., 1993, 1995; Goldstein and Spiegelhalter, 1996; Normand et al., 1997; Thomas et al., 1994).

Normand et al. (1997) presented a framework for hierarchical methods for provider profiling, incorporating complex variation to allow for different hospital types to have different between-hospital variation. They used Bayesian methods to present provider effects as posterior tail probabilities of different measures of poor performance, arguing that these are a more interpretable quantity than the confidence intervals obtained using frequentist methods. Goldstein and Spiegelhalter (1996) also exploited the flexibility of Bayesian models to illuminate flaws in simple league tables because of high levels of uncertainty in the ranks. They highlighted that extension of their methods to three-level models is straightforward, enabling profiling of providers within providers.

Three-level models have been implemented in several studies, for example to examine healthcare-associated infections in hospitals within regions in France (Chen et al., 2013), and adherence to osteoporosis guidelines for physicians within clinics (Brookhart et al., 2006).

In line with focus of public interest described in chapter 1, subsequent research is concentrated in developing methods for detecting unusual performance using techniques such as funnel plots (Jones et al., 2008; Spiegelhalter, 2005b). Similarly, the Standardised Mortality Ratio (SMR, which compares the observed deaths with the number expected) is an intuitive, if much-criticised, measure of hospital performance (Lilford and Pronovost, 2010), but it is less clear within the hierarchical framework how the numerator and denominator should be calculated. Several different definitions have been proposed (Glance et al., 2003; Health and Social Care Information Centre Clinical Indicators Team, 2016; Shahian et al., 2005; Yang et al., 2013) but these do not always yield consistent results (Mohammed et al., 2012). Whether we wish to quantify provider effects using odds ratios, SMRs, ranks or tail probabilities, in the context of hierarchical models all of these approaches involve estimating the random effects themselves in some way. As previously described shrunken estimates of provider effects can be calculated using empirical or full Bayes. The former was used by Thomas et al. (1994) amongst others (DeLong et al., 1997; Dimick et al., 2010) and is sometimes called a reliability adjustment. Gatsonis et al. (1993) found both approaches gave similar results for their purpose.

Provider profiling differs from many standard applications of hierarchical models (e.g. repeated measures, growth curves, cluster randomised trials) in that the random effects themselves are the focus of interest, not a nuisance source of variation to be accounted for in order to correctly estimate other parameters (Ohlssen et al., 2007b). Therefore the random effects should be carefully specified with regard to the inferences we wish to make, with sensitivity analysis to check for robustness to variation in the assumptions. Several studies have explored using alternative distributions to Normal for the random effects. For provider profiling analyses a t distribution with low degrees of freedom has been suggested (Normand et al., 1997; Ohlssen et al., 2007b) but there are no specific recommendations about exact values for ν. Values of $\nu = 1$ or 2 have an infinite variance so are not appropriate choices. Within a Bayesian framework we can attempt to estimate ν from the data (setting a suitable prior) but unless the data (or prior) are very informative this does not yield a useful posterior and is similar to specifying a Normal distribution (Daniels and Gatsonis, 1999). Austin (2005) demonstrated using simulation that assuming a normal model to estimate provider effects arising from a t distribution with $\nu = 3$ and $\nu = 5$ resulted in over-shrinkage but there was no basis for choosing ν. In some situations a skewed distribution may be suitable, for example if there are a few providers who are particularly high performing. Lee and Thompson

(2008) present an example concerning a clinical trial comparing teleconsultations with standard appointments, where treatment effects were expected to differ by consultant. They used flexible Normal and t distributions for the random effects with parameters, estimating both ν and a skewness parameter from the data, though in this case there was little difference in model fit compared with symmetry.

There are some limitations to using hierarchical models for provider profiling. While shrinkage of providers is desired for some purposes it can lead to a false sense of complacency for smaller providers (Walker et al., 2013). In such cases the tail probabilities advocated by Normand et al. (1997) may be useful to indicate the extent to which unusual performance can be ruled out. Another consequence of shrinkage is that smaller providers will be shrunk more towards the mean than larger providers. Direct comparison of posterior means between providers may therefore yield incorrect inferences (Shen and Louis, 1998). Aylin et al. (2003) argue that if an aim of the analysis is to detect outliers then it may be preferable to minimise the risk of type II error (not detecting a genuine outlier) in comparison with the type I error (erroneously identifying a provider as an outlier); in such cases a fixed effects model may be preferable to a hierarchical model. A similar argument was made by Austin et al. (2003) who demonstrated that hierarchical models in general had greater specificity for correctly identifying outliers while fixed effects models had greater sensitivity, even when all providers were high-volume centres. Notably, the authors found that fixed effects models outperformed hierarchical models if the underlying distribution of provider effects was trimodal, but a unimodal random effects distribution was assumed. The requirement to assume a distribution for the random effects can therefore be a limitation, one which is not present for fixed effects models.

Hierarchical models also require the assumption that the provider effects are uncorrelated with covariates. This can be violated if there are unmeasured confounders (i.e. variables which are associated with both the provider and the outcome that are not in the model) that are also associated with included covariates, because the effects of the omitted variables are absorbed into the provider effects. Sociodemographic variables such as ethnicity and economic deprivation can be confounders because they have spatial variability and have effects on health outcomes, but provider profiling analyses are often not adjusted for them to avoid masking the effects of interest. This was discussed in chapter 3, along with the effect on the interpretation that some of the between-provider variation could be due to unmeasured confounding. A further effect of adopting this approach occurs if the unmeasured sociodemographic variables, and hence the provider effects, are correlated with any of the clinical risk-adjustment factors. Then the assumption of independence of provider effects and covariates is violated, resulting in biased estimates of the covariate effect (Rabe-Hesketh and Skrondal, 2005). An advantage of fixed effects models is that they do not require that provider effects be independent from

covariate effects.

The concept of hierarchical modelling (and caveats associated with the results) can be difficult to explain to a non-statistical audience. Sometimes these difficulties relate to aspects which are otherwise considered advantageous. The Centers for Medicare and Medicaid Services (CMS) in the US received several criticisms from consumers regarding the use of hierarchical models (Ash et al., 2012): the methods did not show any variation because of shrinkage and were therefore not useful; the performance of small hospitals is masked; concepts surrounding hierarchical models are difficult to convey. Nevertheless hierarchical models have now gone firmly beyond the statistical literature and have been put into practice by the CMS (Shahian et al., 2005) and many others such as the UK Healthcare Comission (Spiegelhalter, 2005a) (now the Care Quality Commission) and the UK Health and Social Care Information Centre (Health and Social Care Information Centre Clinical Indicators Team, 2016).

In summary, hierarchical models are an established approach for provider profiling, and this thesis builds on these existing approaches. Some aspects of the structure of the neonatal data can be modelled using techniques described thus far. As with all regression models, continuous and categorical patient risk factors can be adjusted for. Hierarchical models allow for variation at both the NNU and network level to be estimated using a three-level model. Level of NNU can be incorporated into the model as a provider-level covariate, and heteroscedasticity by level of NNU investigated by including separate variance terms. These methods will be described in detail in section 5.2. A limitation of these established methods is that they only allow for attribution to a single provider. The problems posed by transfers for provider profiling analyses (see chapter 1) and analysis of transfer patterns (chapter 2) indicated that attribution to a single provider is not a reasonable approach for these data, nor is excluding transfers altogether. In the next section I will review the literature for how transfers have been handled in existing provider profiling studies.

4.2 Literature review of transfers

The objective of this literature review is to find out whether there are any existing studies which either accounted for transfer of patients in hierarchical provider profiling analyses, or considered the theoretical aspects of doing so. Included studies are those which attempted to assign patients to all providers involved in care, but I am also interested in the relative frequency of such studies, compared with those excluding transfers, and those assigning to a single provider.

4.2.1 Search strategy

The search strategy was designed to return papers containing terms relating to hierarchical modelling, provider profiling and transfer. For provider profiling there are few standard phrases which are used consistently across the literature so a variety of search terms were chosen based on re-occurring terminology which featured in initial reading. The exact search terms are shown in table 4.1, with the asterisk indicating a wildcard to allow different forms of the term of interest:

Table 4.1: Results from systematic review

Hierarchical terms	multilevel OR hierarchical OR mixed model OR random effects
Profiling terms	hospital comparison OR institutional comparison
	OR
	(rank* OR profiling OR benchmark*) AND (institut* OR hospital OR provider)
	OR
	(measur* OR compar* OR monitor*) AND (performance OR quality) AND (institut* OR hospital OR provider)
Transfer terms	transfer* OR transport* or discharg*

All terms with multiple words were searched as a single phrase

I have included both the specific phrases for *hosptial/institutional comparison* and the individual words along with *performance/quality*. The rationale is that the phrases alone may not capture relevant results, but including the individual words without *performance/quality* yields many irrelevant results as what is being measured or compared does not relate to providers. Therefore specific phrases are included, along with the individual words with the requirement that *performance/quality* be present as well.

I searched the PubMed database, using Medical Subject Headings (MeSH terms) where available, searching all fields. Titles and abstracts of the resulting papers were reviewed with papers selected for full-text review provided they appeared to concern provider profiling, with the possibility of using hierarchical models as this was not always mentioned in the abstract. On full text review, studies were considered relevant if the treatment of transferred patients in the analysis was mentioned, where a *transferred patient* was any patient treated at more than one provider included in the analysis. Studies which either excluded transfers or assigned to a single provider were noted, and studies assigning to multiple providers were reviewed more fully. The selection was

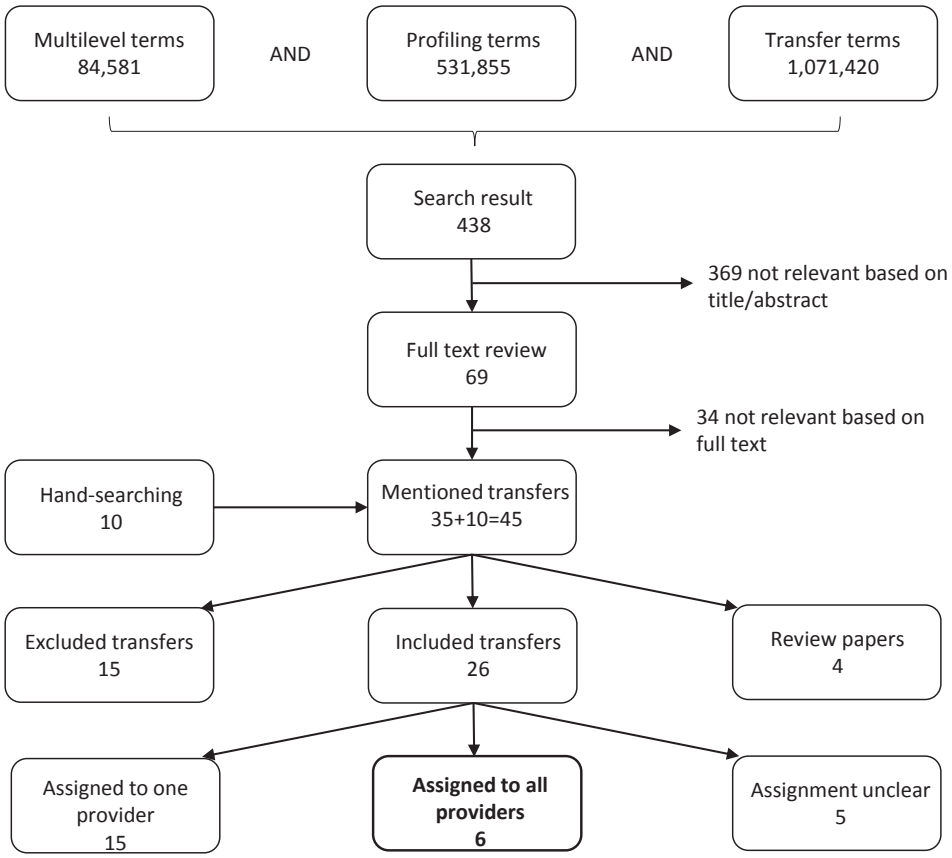


Figure 4.2: Flowchart illustrating the literature review

limited to papers published in English where the full text of the paper was available. Reference lists and forward citations of the relevant texts were hand-searched. Review papers were included where relevant.

4.2.2 Results

The PubMed search yielded 438 results, of which 35 were relevant and a further 10 articles were found from hand-searching reference lists or forward citations of relevant texts, yielding a total of 45 papers which mentioned transfers. Figure 4.2 is a flowchart illustrating the literature review. The search is up to date as of 11th April 2016.

Table 4.2: Results from systematic review *(continues to table 4.3)*

Study	Population	Data	Transferred patients	Clinical outcome	Providers	Performance measure	Treatment of transfers
Kosseim 2006	91,633 adults with a primary diagnosis of AMI 1992-1999	Administrative data (hospital and civil registries)	15% transferred once	30-day mortality	116 acute care hospitals in Quebec, Canada	Spread and rank of risk adjusted mortality rate (the probability of death of an patient with baseline covariate values including the EB estimate of the hospital random effect)	Compared (i) excluding transfers (ii) assign to referring hospital (iii) assign to receiving hospital
Rasmussen 2008	60,339 adults surviving first AMI 1995-2004	Administrative data (hospital, civil and pharmacy regisitries).	Not stated	Prescription claim for preventative medicine within 90 days discharge	73 acute care hospitals in Denmark	80% prediction range of hospital effects calculated from hospital variance parameters. Hospital effects were modelled to vary over time. Multivariate normal distribution for random effects, method of estimating random effects not stated.	Transfers were assigned to hospital of first admission, but assigned to discharge hospital in a sensitivty analysis.
Campbell 2012	36.5 million hospital admissions 2005-2010	Administrative (HES and death registries)	<1% of admissions were across multiple hospitals	Mortality in hospital or within 30 days of discharge	146 general and 72 specialist hospital trusts in England	SHMI, similar to a SMR. These were not calculated using multilevel models, though a random effects analysis was used to determine outliers.	Admissions were the unit of analysis. If a patient was admitted to several hospitals then died, the death was attributed to the last hospital, and the preceeding admission treated as survival.
Bottle 2013	329,369 ACS admissions in 288,550 patients Apr 2006-Mar 2009	Administrative (HES and death registries)	11.7% patients with more than one admission, with a maximum of 10 admissions.	Mortality in hospital or within 30 days of discharge	146 acute, non-specialist hospitals in England.	SMR, calculated as the observed deaths divided by the expected deaths from the fixed part of a hierarchical regression model; funnel plot outliers though method for control limits not stated. Distribution and estimation method for random effects not stated.	Compared i) include all transferred patients and adjust for whether they were transferred in; ii) exclude transfers in; iii) exclude patients transferred in or out; and iv) count transfers as separate admissions, and compared including comorbidity data from previous admissions.

AMI Acute Myocardial Infarction; EB Empirical Bayes; HES Hospital Episode Statistics; SHMI Summary Hospital Mortality Index; SMR Standardised Mortality Ratio; ACS Acute Coronary Syndrome

Table 4.3: Results from systematic review *(continued from table 4.2)*

Study	Population	Data	Transferred patients	Clinical outcome	Providers	Performance measure	Treatment of transfers
Samadashvili 2015	81,852 AMI admissions Jan 2008-Dec 2010	Administrative (hospital and social security registries)	22.4% overall, ranging from 1.3% to 54.4% depending on hospital type	Mortality in hospital or within 30 days of discharge	194 acute care hopsitals in New York State	RMSR = predicted/expected mortality rates, multiplied by statewide rate, no further details of calculation. Outliers were those where statewide rate lay outside bootstrap confidence interval for the hospital RSMR.	Compared attribution to transferrring and receiving hospital for emergency department transfers.
Hassani 2015	Patients admitted for first time AMI, stroke or hip fracture 2011-2013. Number not stated.	Administrative (hospital and civil registries)	Up to 50% for AMI.	Mortality in hospital or within 30 days of discharge	52 hospitals in 22 hospital trusts in 4 regional health authorities in Norway.	Hospital effects obtained using FE. Outliers determined based the trimmed hospital effects distribution applying a multiple testing correction. Estimated survival rate amongst the whole cohort if they had attended each hospital = sum of survival probabilities replacing hospital effects with a shrunken hospital effect for the hospital in question. The FE were shrunk using a hierachical Bayesian model assuming a t-5 prior for the hospital effects.	Weighted fixed effects proportional to length of stay, only counting time on the ward.

AMI Acute Myocardial Infarction; RSMR Risk-Standardised Mortality Rate; FE Fixed effects

These 45 papers were applications or review articles concerning provider profiling using hierarchical models which mentioned the treatment of transferred patients (see appendix C for the full list of studies). Four of these (Bruckel et al., 2016; Jollis and Romano, 1998; Krumholz et al., 2007, 2006) were review papers which discussed transfers but none provided suggestions for allocation to multiple providers.

Of the 15 studies which excluded transfers, only four of those (Aelvoet et al., 2016; Reeves et al., 2010; Reistetter et al., 2015; Sethi et al., 2007) mentioned the number of patients that this applied to, with proportions ranging from 2% to 37%. Another 15 studies assigned transfers to a single provider, of which eight studies (Ben-Josef et al., 2014; Bratzler et al., 2011; Kim et al., 2015; Marcin and Romano, 2004; Reineck et al., 2014; Rumsfeld et al., 2006; Seghieri et al., 2012; Suter et al., 2014), allocated patients to the first provider, in line with recommendations from three of the review papers (Jollis and Romano, 1998; Krumholz et al., 2007, 2006). The rationale generally presented was that the first provider is responsible for the decision to transfer, and allocation to the receiving provider inappropriately penalizes those receiving more complex cases (assuming this is not captured by risk-adjustment), as well as potentially encouraging premature transfer to avoid attribution of a negative outcome. Five studies (Dégano et al., 2015; Gomez et al., 2015; Jan et al., 2013; White et al., 2013; Xu et al., 2015) mentioned that transfers were included but did not state how they were assigned.

Therefore of a total of 41 studies (excluding review papers) of hierarchical provider profiling analyses which mention transferred patients, only six (Bottle et al., 2013; Campbell et al., 2012; Hassani et al., 2015; Kosseim et al., 2006; Rasmussen and Abildstrom, 2008; Samadashvili et al., 2015) attempted to assign patients to multiple providers. Study summaries of these six papers are shown in tables 4.2 and 4.3. Three studies Bottle et al. (2013); Kosseim et al. (2006); Samadashvili et al. (2015) specifically set out to investigate the impact of transferred patients on variations in mortality across providers. Bottle et al. (2013) were concerned with the impact of excluding transferred patients, while Kosseim et al. (2006) and Samadashvili et al. (2015) sought to compare allocation to the transferring and receiving hospital. Campbell et al. (2012) and Hassani et al. (2015) documented methods used for national comparisons for hospital mortality in England and Norway respectively, while Rasmussen and Abildstrom (2008) conducted a research study examining trends in medicine use.

Setting and outcomes All but one study concerned mortality in adults following hospital admission for acute coronary syndrome (ACS) or acute myocardial infarction (AMI). The proportion of ACS/AMI patients undergoing transfer ranged from 1% to over 50%; transfer rates were much lower in the remaining populations. Only one study (Bottle et al., 2013) mentioned the number of admissions per patient (at most ten) but

in two studies it was implied by the ability to assign patients to either the transferring or receiving hospital that patients were only transferred once (Kosseim et al., 2006; Samadashvili et al., 2015). All studies used administrative databases rather than medical records as the data source for analysis. As Bottle et al. (2013) remark, administrative databases are often used for provider profiling analyses, and they often contain limited information which may make it difficult to link episodes of transferred patients. Three studies (Hassani et al., 2015; Kosseim et al., 2006; Samadashvili et al., 2015) calculated a risk-standardised mortality rate, though all were calculated in different ways. Two studies (Bottle et al., 2013; Campbell et al., 2012) calculated standardised mortality ratios (again using slightly different methods) and the remaining study (Rasmussen and Abildstrom, 2008) considered variation in the odds of death using odds ratios. Most of the studies were interested in best and worst performers, outliers and overall variation measured by ranges, with no studies examining changes in provider effects for individual hospitals.

Methods Four studies (Bottle et al., 2013; Kosseim et al., 2006; Rasmussen and Abildstrom, 2008; Samadashvili et al., 2015) conducted multiple separate analyses, each of which assigned the patient to a single provider. Two of these (Bottle et al., 2013; Kosseim et al., 2006) in addition compared results with the frequently used strategy of excluding transferred patients. Bottle et al. (2013) also conducted an analysis treating admissions as the unit of analysis; a similar strategy was taken by Campbell et al. (2012). In these analyses, if the patient was admitted to several hospitals and then died, the death was attributed to the last hospital and the preceding admission(s) treated as survivals. This counts transferred patients multiple times, but without accounting for the dependence between these outcomes (as patients can only die once). However in Campbell et al. (2012) less than 1% of patients were transferred and in Bottle et al. (2013) this was only one of four analyses and was designed to illustrate the effect of ignoring all transfer information.

Hassani et al. (2015) was the only study which attempted to allocate patients to all providers without considering each admission separately. In this study, a fixed effects analysis was used including a parameter for each hospital treating the patient, weighted according to the proportion of their total stay which took place in the hospital in question. The study determined outliers using these fixed effects regression coefficients, but then applied a Bayesian hierarchical model to shrink the coefficients assuming they follow a t distribution with 5 degrees of freedom, applying these shrunken coefficients when calculating a risk adjusted mortality rate. The two-stage nature of this analysis means that any uncertainty from the first stage is not carried through. Hassani et al. (2015) did not discuss transfers or their approach to accommodating them in detail, but they cite another study (Kristoffersen et al., 2012) from the same group for their method.

Kristoffersen et al. (2012) also used weighted fixed effects, but did not use hierarchical modelling and so did not meet the search criteria for the review. They compared weighting according to length of stay with excluding transfers and treating each admission separately. They applied weights proportional to length of stay arguing that this method has simplicity and transparency in its favour, but suggested weighting critical care more heavily if such information is available. The authors of the study were not aware of any previous study attempt to assign transferred patients to multiple hospitals.

Results Several authors found that results were sensitive to how patients were allocated, with no clear preferred way of attributing transferred patients (Kosseim et al., 2006; Kristoffersen et al., 2012; Samadashvili et al., 2015). Of the three studies which sought to compare results from different ways of allocating transfers, two found differences in which hospitals were the top and bottom performing providers (Kristoffersen et al., 2012; Samadashvili et al., 2015), though neither study shows whether these differences are notable compared with random variation. Kristoffersen et al. (2012) also found substantial differences in ranks and outlier status. However no adjustment was made for multiple testing, no measures of uncertainty around the ranks were calculated and outliers were determined simply by the extremes of the hospital effects, which will exist even if all hospitals are performing equally well.

Kosseim et al. (2006) suggest that if results for a hospital improve when patients are assigned to the referring hospital it suggests that the hospital is doing well at stabilising them prior to transfer. However it could instead be a reflection of the good performance at the receiving hospital being erroneously attributed. Kosseim et al. (2006) and Samadashvili et al. (2015) both found that patients who were transferred were generally less sick (lower mortality, younger and with fewer co-morbidities). Samadashvili et al. (2015) was the only study to look at transfers by hospital type, finding differences in rates of transfer, and a mortality difference in transferred and non-transferred patients.

Bottle et al. (2013) also investigated how methods for handling transfers influenced results, and found minimal differences in the SMR between the models including transfers, but notable differences if all transfers were excluded. Kosseim et al. (2006) found the variation in risk adjusted rates was similar whichever method was used to allocate transfers. Rasmussen and Abildstrom (2008) stated that their findings (one of which was between hospital variation) were robust to allocating transfers to the discharge hospital rather than first admission, but no further details are provided.

The exact methods, particularly regarding the specification and estimation of hierarchical models were unclear in several of the studies (Bottle et al., 2013; Rasmussen and Abildstrom, 2008; Samadashvili et al., 2015) and data dependencies were ignored in other ways (multiple episodes from the same patient in Bottle et al. (2013) and Camp-

bell et al. (2012), and higher levels of hierarchy in Hassani et al. (2015)). Several studies did not fully incorporate uncertainty into findings, looking at measures such as ranks and tertiles without a measure of uncertainty (Kosseim et al., 2006; Samadashvili et al., 2015) or using fixed effects analysis without adjusting for multiple testing (Bottle et al., 2013).

Discussion of transfers Some authors cited centralisation (only some hospitals performing more specialised procedures) as a reason for transfers becoming more common (Kosseim et al., 2006; Kristoffersen et al., 2012; Samadashvili et al., 2015). Kristoffersen et al. (2012) also noted the increase in transfers due to centralisation, but found frequent transfers from larger to smaller hospitals for patients admitted for stroke and hip fracture, representing rehabilitatory care.

Several studies commented on the difficulties that transfers pose. Two studies (Campbell et al., 2012; Kosseim et al., 2006) highlighted that patient transfers can be a product of clinical care (if the care provided affects the condition of a patient, determining clinical need for transfer) and management (the decision-making and logistics behind the transfer itself), with delayed transfer and transfer of unstable patients both leading to poorer outcomes. Bottle et al. (2013) and Rasmussen and Abildstrom (2008) both adjusted for transfer status (the former in some of the comparative analyses) but neither of these studies considered that transfers could be a reflection of care provided. Though this issue was not specifically addressed by the study, Campbell et al. (2012) highlighted that all providers treating a patient can influence the outcome so it is not strictly fair to assign their outcome to one provider alone. They applied the advice from the UK Department of Health (who commissioned and funded the study) in allocating deaths to the final provider but pointed out (as did Bottle et al. (2013) and Samadashvili et al. (2015)) that allocation to the receiving hospital could lead to premature, inappropriate or unnecessary transfers in order to "game" the system, resulting in biased hospital effects and, more seriously, adverse consequences for patients. This is the reason that it is usually recommended to assign transfers to the referring hospital (Jollis and Romano, 1998; Krumholz et al., 2007, 2006; Samadashvili et al., 2015). However it is done, allocation to a single provider means that we are attributing the performance of one provider to another. Campbell et al. (2012) (citing methodology used by Dr Foster (Aylin et al., 2010), who provide independent assessments of NHS hospitals) suggest that allocating outcomes to all providers is possible in some situations if outcomes are calculated separately for each hospital rather than using a hierarchical model.

The rationale for assigning patients to multiple providers was generally practical, concerning sensitivity of results and fairness in allocation rather than use of statistically appropriate methodology. None of the studies mentioned that by assigning individuals

to a single provider, data dependencies are not correctly accounted and the estimates of between-hospital variation may be incorrect .

In this review several studies have considered the problem of transfers. The studies indicated scope for different methods outside the hierarchical modelling context using fixed effects (Hassani et al., 2015; Kristoffersen et al., 2012), allocating to all hospitals involved by estimating hospital effects outside the model (Aylin et al., 2010; Campbell et al., 2012) or by treating each episode separately (Bottle et al., 2013; Campbell et al., 2012). However there were no satisfactory solutions using hierarchical models. In the next section I propose using multiple membership models as a way of including transferred patients in hierarchical models.

4.3 Multiple membership models

The hierarchical models introduced in section 4.1 can be extended to allow individuals to belong to more than one cluster, e.g. for patients treated at more than one provider. These models are known as *multiple membership* (MM) models (Browne et al., 2001; Fielding and Goldstein, 2006; Goldstein, 2002). The MM structure for data with two levels is illustrated in figure 4.3(a) in contrast to the single membership structure in 4.1(a). The aim of this section is to introduce multiple membership models and describe some of the existing literature.

4.3.1 Definition

In the basic hierarchical model described in equation 4.2 for an individual i in a single cluster j, we had one random effect u_j representing the effect of cluster j on the log odds of the outcome Y. The multiple membership is modelled by replacing this single effect with a weighted sum of random effects, one for each cluster of which the individual is a member, with the weights constrained to sum to one for each individual. This is similar to the method implemented by Hassani et al. (2015) in a fixed effects context. A change in notation is required because of the multiple membership: we drop the j subscript for the individual as they are no longer a member of a single cluster j and let $U(i)$ denote the set of clusters to which individual i belongs (sometimes referred to as *classification* notation (Goldstein, 2002)). Extending the two-level random effects model in equation 4.2 we have:

$$\begin{aligned} logit(p_i) &= \alpha + \beta x_i + \sum\nolimits_{j \in \mathrm{U}(i)} w_{ij} u_j \\ u_j &\sim N(0, \sigma_u^2) \qquad \sum\nolimits_{j \in \mathrm{U(i)}} w_{ij} = 1 \end{aligned} \tag{4.5}$$

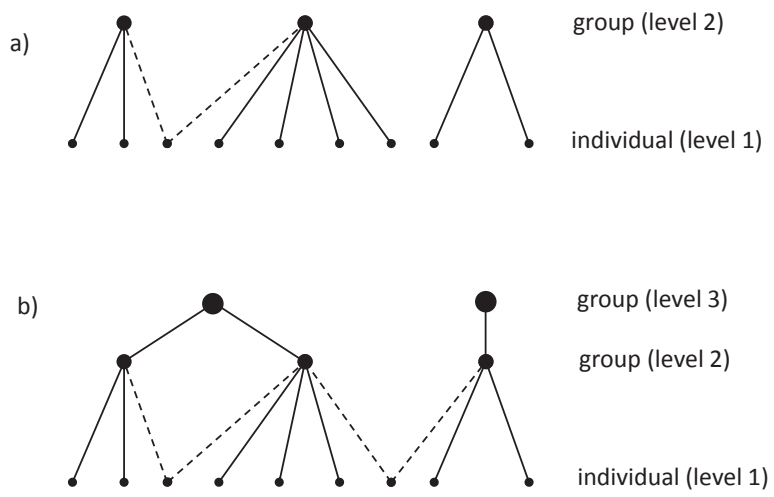


Figure 4.3: Diagrams illustrating multiple membership (MM) data structures: (a) as figure 4.1(b) but with one individual being a member of two groups; MM indicated by dashed lines) (b) as figure 4.3(b) but with two individuals with MM, indicated by dashed lines)

The choice of weights will be discussed in section 4.3.2. As with other hierarchical models, the random effects u_j are often assumed to have a Normal distribution, but the same considerations discussed in section 4.1 apply. Likelihood-based methods for standard hierarchical models exploit the nested structure of the data which is not present for MM models (Goldstein et al., 2007). MM models are more straightforward to estimate using Bayesian Markov Chain Monte Carlo (MCMC) estimation because of the flexibility of the iterative approach (Browne et al., 2001); these methods will be discussed in section 5.1.

Any cluster-level variables need to be weighted in the same way as the random effects (Fielding and Yang, 2005; Goldstein et al., 2007). For this reason the hierarchical centring parametrisation is advantageous as we avoid specifying multiple terms for each cluster at the individual level. We can also have MM occurring in three-level models. Figure 4.3(b) shows MM only at the individual level, but it is also possible for level 2 clusters to belong to more than one level 3 cluster. Let $V(j)$ denote the set of level 3 clusters to which level 2 cluster j belongs. We can then extend the three-level model described in 4.3 to incorporate multiple membership, and also add a continuous covariate

Z at level 2:

$$
\begin{aligned}
logit(p_i) &= \beta x_i + \sum\nolimits_{j \in \mathrm{U}(i)} w_{ij}\mu_j && \text{Level 1} \\
\mu_j &= \gamma z_j + \sum\nolimits_{k \in \mathrm{V}(j)} w_{jk}\nu_k + u_j && \text{Level 2} \\
\nu_k &= \alpha + v_k && \text{Level 3} \\
u_j &\sim N(0, \sigma_u^2) \\
v_k &\sim N(0, \sigma_v^2)
\end{aligned}
\tag{4.6}
$$

4.3.2 Existing literature on multiple membership models

Multiple membership models were first introduced by Hill and Goldstein (1998) as a way of modelling uncertainty about which higher level unit an individual belongs to. In their example, the weight w_{ij} represented the probability that a pupil i attended school j, where the probabilities could be informed by data on location of school, and previous or subsequent schooling of the pupil. Several studies (Cafri et al., 2015; Goldstein et al., 2007; Leckie, 2009) have used MM models to examine pupil mobility, finding notable differences compared with single membership models (those allocating pupils to only one of the schools attended). Leckie (2009) noted that, in contrast to the hospital transfers studies described in section 4.2, educational studies tended to assign pupils to the final school attended. In the medical literature, Roberts and Walwyn (2012) used MM models to allow for therapist effects in psychotherapy trials where patients are treated by more than one therapist. Life course epidemiology studies have used MM models as a way of modelling spatially and temporally correlated health effects (Ali et al., 2007; Chandola et al., 2005; Leyland and Næss, 2009). Leyland and Næss (2009) modelled the effect of area of residence at three different census points on subsequent mortality, allowing for complex variation by time point to reflect the varying strength of influence over time. Silcocks and Kendrick (2010) also applied MM models to spatial effects in the context of overlapping populations in cluster randomised trials. Where data are longitudinal with repeated measurements, and cluster membership available at a series of time points, MM models can be used to estimate carryover effects (e.g. a primary school's influence on a pupil's test result in the first year of secondary school) (Cafri et al., 2015).

Comparison with single membership models It has been shown that ignoring MM structures can lead to biased estimates of both fixed and random parameters (Browne et al., 2001; Chung and Beretvas, 2012). Ignoring the MM structure and simply assigning individuals to one cluster has been shown to result in underestimation of the cluster-level variance (Chung and Beretvas, 2012; Fielding and Goldstein, 2006; Goldstein et al., 2007; Leckie, 2009). Chung and Beretvas (2012), when examining the

impact of ignoring MM structure using simulated data, found an average 16% downward bias in the between-cluster variance, with the level of bias dependent on the proportion of individuals with multiple membership. The contribution of the clusters to the total variation is, for an individual i:

$$var(\sum\nolimits_{j\in \text{unit}(i)} w_{ij}u_j) = \sigma_u^2 \sum\nolimits_{j\in \text{unit}(i)} w_{ij}^2 \tag{4.7}$$

Under single membership this is simply σ_u^2, but for an individual that is a member of more than one cluster, the sum of w_{ij} is always less than 1, so the variance contribution is less than σ_u^2. Using the example of pupil mobility and educational outcomes described by Goldstein (2002) and Leckie (2009), suppose a pupil attends two schools for an equal period of time. For this pupil, the contribution of schools to the total variance is $0.5^2\sigma_u^2 + 0.5^2\sigma_u^2 = 0.5\sigma_u^2$, so the observed variance for this individual is only half the true school-level variance. Ignoring the MM structure implicitly assumes the school-level variance contribution is the same for all individuals, regardless of pupil mobility, thereby underestimating the between-school variance and producing estimated school effects $\hat{u}_j$ which biased towards zero. The relative importance of schools in the overall variation in educational outcome (as measured by the intraclass correlation coefficient or variance partitioning coefficient, which will be described in section 5.2.6) will also be reduced. As demonstrated by Chung and Beretvas (2012), as more pupils move schools, the variance underestimation increases. For a pupil attending four schools, the variance attributable to schools is $4 \times (0.25^2\sigma_u^2) = 0.25\sigma_u^2$. This has an intuitive interpretation in that the more schools a pupil attends, the more "average" their experience is. If a single membership model is assumed, the variance underestimation will more marked in situations where pupils are highly mobile.

Choice of weights For membership defined by time (e.g. how many years a pupil attends a school) weights could be defined as the proportion of time spent in each cluster; this is a standard choice (Browne et al., 2001; Fielding and Goldstein, 2006; Goldstein et al., 2007; Leckie, 2009). For membership defined by space (e.g. the neighbourhood effects modelled by Silcocks and Kendrick (2010)) weights could be defined by the relative distance from the individual. For MM models involving discrete time periods, it is possible to use fixed weights. For example, secondary education from 11 to 16 years of age, if the last two years representing GCSE examinations were considered most important, weights could be set to assign 10% each to the first three years and 35% each to the final two years. An alternative parametric approach was suggested by Leyland and Næss (2009) where the weights are proportional to the square root of the between-cluster variance at that time point. However this is only possible when each cluster membership is defined for fixed time points (in that study, area of residence for three different census years) so that a separate variance can be estimated for each cluster effects at the same

time point.

The analyst has to make a choice regarding the weighting scheme to be used. In equation 4.7, we see that individuals with different MM patterns will have different amounts of their variance that is attributable to clustering. Similarly to the bias when estimating σ_u^2 and u_j that occurs if the single membership is assumed, we might expect biased results if we apply a MM weighting scheme that does not reflect the data. In practice, studies using MM models to evaluate effects of pupil mobility found that results were robust to alternative weightings (Goldstein et al., 2006; Leckie, 2009), as did a study of teachers in multiple teaching groups (Fielding and Yang, 2005). Goldstein et al. (2006) applied a variety of weightings placing different relative importance on different years of schooling as a sensitivity analysis, but found little difference and selected the final model based on Deviance Information Criterion, a Bayesian measure of model fit which will be defined in section 5.1. A similar approach of comparing several weights was taken by (Silcocks and Kendrick, 2010) using Akaike Information Criterion (see 3.4.1 for definition). Wolff-Smith and Beretvas (2014) investigated the effects of misspecifying weights using simulation. The authors generated data where the effect of school attended had an equal influence on the outcome in some datasets and unequal in others. They then analysed the data under the true model and a variety of assumed models. They found that cluster variance was underestimated when unequal weights were used to analyse the equal weight data. Returning to the simple example in the previous section, suppose we analyse the pupil mobility data with a MM model that places 90% of the weight on the first school. The pupil who attends two schools for an equal period of time will now be assumed to have a school-level variance contribution of $0.9^2\sigma_u^2 + 0.1^2\sigma_u^2 = 0.82\sigma_u^2$. If the observed variance for the pupil is $0.5\sigma_u^2$ as before, then this leads to an underestimation of the school-level variance. We might expect the converse (overestimation of the σ_u^2 when using equal weights to analyse unequal data) as the variances for individuals will be assumed to be smaller fraction of σ_u^2 than they really are. Wolff-Smith and Beretvas (2014) found this to be the case only in two scenarios, in which applying the true weighting also overestimated the variance, so this may be an artefact of the data generation process. In this study, as with the others mentioned here, estimation of cluster-specific results as not the primary aim with analysis limited to comparing cluster-level variance and the correlating the ranks of cluster effects.

4.4 Discussion

4.4.1 Summary

In this chapter I reviewed existing statistical methods relevant to the research questions of the thesis. I introduced hierarchical models and described their applicability to provider profiling studies in general and for this thesis. I conducted a systematic review of the literature to ascertain existing strategies for accommodating transferred patients using hierarchical models. Almost all studies including transferred patients allocated them to a single provider, or conducted separate analyses comparing different single allocations. Only one study attempted to allocate patients to all providers but this was done using a fixed effects model, with a hierarchical analysis applied afterwards to shrink the provider effects. There were four review papers discussing the problem of transferred patients but none presented a suitable solution. The results of the systematic review highlighted a gap in knowledge. I then introduced multiple membership models which weight the random effects as a possible solution to the problem of transfers. Multiple membership models were developed in education primarily to model pupil mobility. I described some of the applications and in the literature, and some of the theoretical findings. In this thesis I am particularly interested estimation of the provider effects themselves, but there has been limited attention in the literature as to how these are affected by different weighting choices.

4.4.2 Hierachical models

Hierarchical models are an established and recommended method for provider profiling studies, and are particularly applicable for the research questions of this thesis. Elements of the structure of neonatal services can be reflected in the structure of the hierarchical model. The ability to model clustering at more than two levels is essential because NNUs are nested within neonatal network; ssee Chen et al. (2013) and Brookhart et al. (2006) for implementation of these methods in a provider profiling context. This precludes the use of other methods to handle clustered data such as generalised estimating equations (a method for estimating population-averaged models) as these only accommodate clustering at one level. Allowing the mean and variance of provider effects to be specific to type of NNU follows from the large differences in patterns of care documented in chapter 2, and may be a way of adjusting for differences in case-mix not captured by the risk factors (see chapter 3). This can be incorporated into hierarchical models as set out Normand et al. (1997). The aim of including these aspects in the model is to produce NNU effects which correspond to the inferences we wish to draw as well as to account for clustering in a statistically principled manner. For example including clustering at

the neonatal network level will mean that network effects can be estimated, NNU effects are adjusted for any variation attributable to the network level, and more information is borrowed from other NNUs within the network. In chapter 5 I will propose a selection of models to reflect the different inferences we might wish to make.

Birth before 32 weeks gestation is relatively rare and subsequent mortality rarer still, with a wide range in the volume of care provided at NNUs (see section 2.3.2). There are some NNUs with very small numbers, in particular special care baby units (SCBUs)[1] perform very little care for infants at or below 32 weeks gestation, with a median of 47 infants admitted across SCBUs over the three year period. Therefore shrunken estimates of NNU effects as obtained by emprical Bayes or full Bayesian methods are desirable as they will be more reliable for smaller providers. The variation in volume means that we are likely to have different amounts of shrinkage (Rabe-Hesketh and Skrondal, 2005) (with results for smaller NNUs being shrunk closer to the mean). Using NNU effects from hierarchical models to detect outliers or directly compare individual NNUs (e.g. using ranks) is therefore inadvisable, but they are useful from the perspective of an individual NNU. The NNU effects tell us how the expected mortality for infants in their care compares with that of similar infants attending similar NNUs (with the extent of the similarities depending on the model). This is the clinical purpose of the provider profiling methods developed in this thesis. To this end, in chapter 8 I present a case study for a single NNU as an example of how the findings from this thesis could be presented for clinical purposes.

Alternatives to the hierarchical approach would be to estimate NNU effects independently using a fixed effects model, or to calculate Standardised Mortality Ratios (SMRs). While there are potential solutions to the problem of transferred patients using fixed effects models (Hassani et al., 2015) and SMRs (Aylin et al., 2010). I will not apply them here. The large number of NNUs renders the fixed effects approach infeasible, as well as leading to effects being poorly estimated due to low volumes. Fixed effects models cannot model variation by NNU level, or clustering by neonatal network, both of which are central to the investigations of this thesis. Similar problems occur for the standard definition of the SMR using a fixed effect approach. Definitions using hierarchical models have been proposed, but they give inconsistent results (Mohammed et al., 2012). Therefore to avoid ambiguity in interpretation I will express NNU effects as adjusted odds ratios obtained from exponentiating the random effects.

A limitation to using hierarchical models is that they require that the random effects to be uncorrelated with covariates. For the analysis in this thesis, this would occur if NNU effects are correlated with the following clinical risk factors included in the

[1] one of the three categories of neonatal unit as described in section 1.2.1, the others being local neonatal unit (LNU) and neonatal intensive care unit (NICU)

model: gestational age, birth weight, infant sex and whether or not the mother received antenatal steroids. I have chosen not to adjust for any sociodemographic variables based on evidence for association, data quality and resulting interpretation of provider effects (see sections 3.1.1 and 3.2). Any residual confounding effects of these variables will be absorbed into the NNU effects, so a correlation can be induced if there is an association between the sociodemographic variables and clinical risk factors and the former vary by NNU. Gestational age at birth exhibits variation by ethnicity; for example, black mothers are likely to deliver earlier (Goldenberg et al., 2008). As Kramer et al. (2001) report, while socioeconomic status is unlikely to have a direct effect, it is associated with stress and health behaviours that can lead to prematurity and lower birth weight. Both ethnicity and socioeconomic deprivation vary by NNU, and ethnicity was found to be a frequent independent predictor of mortality in preterm infants (Medlock et al., 2011) though the evidence for the effects of SES on mortality beyond that mediated by gestation are unclear (see section 3.2.2). Therefore a limitation of this thesis is that this assumption may be violated, which would lead to biased estimates of the covariate effect (Rabe-Hesketh and Skrondal, 2005), as well as higher between-NNU variance from the inclusion of unmeasured confounders that was acknowledged in chapter 3. This could be avoided by using a fixed effect or population-averaged models, but discussed earlier in this section they are unsuitable for the purposes of this thesis. Stratification by ethnicity and SES would reduce the effects of unmeasured confounding but is problematic because of data quality and the number of infants at each NNU. Another possible solution to address the bias in the covariate effects for gestation and birth weight is to include calculate the mean covariate value at each NNU and include this in the regression model (Bafumi and Gelman, 2006; Mundlak, 1978; Rabe-Hesketh and Skrondal, 2005). However the NNU effect would then be interpreted as comparing mortality with what would be expected for an NNU with the same mean gestational age and birth weight, as well as NNU level, neonatal network and infant-level clinical characteristics. Exploration of this issue, including the impact on results and possible solutions, remains as further work beyond this thesis.

Mortality of preterm infants has been analysed using the NNRD by neonatal network as part of annual reports by the Neonatal Data Analysis Unit (2015). The same infant risk factors discussed in chapter 3 were adjusted for, but in other respects the strategy adopted in that analysis differs to the approach taken in this thesis. Infants were assigned to the neonatal network where mothers booked to deliver so the effect of transfers was not considered. Hierarchical models were not used as there were only 23 networks with a reasonable number of infants in each network; instead SMRs were calculated separately for each neonatal network and illustrated using funnel plots as one aim of the analysis was to detect outliers.

Also using the NNRD, the National Neonatal Audit Programme (NNAP) (Royal

College of Paediatrics and Child Health, 2015) assesses adherence to standards for processes and outcomes (which do not currently include mortality) for NNUs across England and Wales. For some audit questions funnel plots were used to identify outliers, with the limits determined from hierarchical models to allow for underlying variation between NNUs. Smaller NNUs (those with fewer than ten eligible infants) were excluded from NNAP outlier analyses; in addition lack of power was less of a problem as the audit questions concerned continuous or more frequent outcomes. For all questions transferred infants were excluded from the analysis, network-level variation was not modelled and types of NNU were not differentiated. In this thesis I am not looking at outlier detection because of insufficient power and variability across NNUs in the level of shrinkage.

There have been several studies in a neonatal care setting which applied hierarchical models to examine variations in clinical outcomes (Bonet et al., 2011; Lee et al., 2013; MacNab et al., 2005). Bonet et al. (2011) examined variation in breastfeeding rates amongst very preterm infants in different European regions and infants were assigned to the NNU where they spent the first 48 hours. Lee et al. (2013) modelled length of stay in Californian NICUs, excluding infants who were transferred out but investigating whether this was related to length of stay. Most similar to this thesis, MacNab et al. (2005) used a hierarchical Bayesian to estimate NNU-specific mortality rates in Canadian NICUs. The study only considered NICUs and adjusted for whether the infant was born in the same hospital but transfers were not mentioned. Therefore the main questions of this thesis regarding networked care were not addressed in this study. Using random coefficient models they examined whether the effects of patients covariates differed across NICUs, finding variation in the effects of being small for gestational age and having a congenital anomaly. In this thesis I am focussing on how to incorporate the structure of neonatal services into a provider profiling analysis so further exploration of risk factor effects are not of immediate concern. However random coefficient models could be a useful extension to identify if a provider has aberrant outcomes for particular patient groups.

4.4.3 Transfers

The problems posed by accounting for transferred patients in provider profiling studies has been raised in the literature, as shown by the review in section 4.2, with several studies recommending allocation to the referring provider (Jollis and Romano, 1998; Krumholz et al., 2007, 2006). This is not possible in neonatal care as infants may have multiple transfers, in which case an NNU can both receive and refer the same infant. This could be interpreted as allocation to the first NNU caring for the infant, but as shown in chapter 2 many infants are transferred within 24 hours of birth. Most surviving infants will remain in neonatal care until they reach term equivalent age (eight weeks for

an infant born at 32 weeks gestation). While initial care is important, it does not seem reasonable to assign their outcome to an NNU caring for them for such a small fraction of total stay, nor does it result in desirable inferences if we are interested in more than initial care. Many studies excluded transfers but as 37% of the cohort were transferred at least once and transferred infants cannot be considered representative of the whole cohort (section 2.3.3), doing so would introduce bias. Furthermore, as discussed in section 3.2.3, several studies mentioned that transfers are an important part of the care provided so patients undergoing transfer should be included. For neonatal care, whether an infant is transferred is not always determinable at birth, and can be affected by events that occur while on the neonatal unit, so excluding this group would mean excluding based on care provided. Several studies attributed increases in transfers to centralisation, demonstrating that the questions surrounding structure of services which this thesis attempts to address are not unique to neonatal care. Samadashvili et al. (2015) examined how results changed for different types of hospital depending on how transfers were allocated; I will take a similar approach for types of NNU.

For reasons outlined earlier in this discussion, hierarchical models are more appropriate for the neonatal data than assuming independence of provider effects. Therefore ways of handling transfers which were revealed in the review (the weighted fixed effects approach of Hassani et al. (2015) and Kristoffersen et al. (2012), and allocation to all provider as done by Aylin et al. (2010)) are not suitable for the neonatal data. While the studies considering admissions as the lowest level of analysis did not properly account for the dependency between episodes from the same patients, there are methods that can do so. Discrete time logistic regression is used to model the conditional probability of an event occurring in a time period given that it has not already occurred. These models would also allow risk factors to vary over the course of postnatal age. This could help more accurately describe the condition of an infant that is transferred in. However, when reviewing risk factors for preterm mortality Medlock et al. (2011) found that the addition of later data did not provide improved prediction over models limited to data available at birth. A hierarchical discrete time logistic model was used by Diya et al. (2010) to examine associations between nurse staffing factors and mortality for elective cardiac surgery patients in Belgium. In this study there were only 4 time periods which all surviving patients progress through. This is not the case for preterm infants: the number of episodes ranges from one to twelve, with varying admission and discharge times and no routine progression, so I have not considered this method further here.

A limitation of the review is that the search strategy may not have captured all relevant studies. The neonatal length of stay study by Lee et al. (2013) is one example as, unlike the other neonatal applications of hierarchical models mentioned previously, transfers were specifically mentioned. This is likely to be because there are many different phrases associated with provider profiling studies (Lee et al. (2013) discussed

variation across NICUs). I did consider expanding the search but many of the terms are very general and would retrieve a large number of results. I mitigated the limitations of the search this by also searching reference lists and forward citations for relevant papers which yielded eleven studies of interest (the ten in figure 4.2 plus Kristoffersen et al. (2012)).

4.4.4 Multiple membership models

Multiple membership (MM) models present an attractive potential solution to accommodating neonatal transfers within the hierarchical framework. For the purposes of this thesis we only consider MM at the individual level (and not a level two group being a member of more than one level three group) because NNUs are only in one network. However, even in this case it is still possible for an individual to be a member of more than one level three group (as illustrated in figure 4.3(b)) describing cross-network transfers. Assigning weights to all clusters an individual belongs to is similar to the approach of Hassani et al. (2015) and Kristoffersen et al. (2012) but within a single hierarchical model. Hassani et al. (2015) conducted a two stage analysis to shrink the provider effects, but this is more cumbersome and does not properly carry through the uncertainty from the first stage into the final model. Multiple membership has been used to model pupil mobility (Goldstein et al., 2007; Leckie, 2009) which is similar in nature to neonatal transfers. The school effects were themselves of interest just as NNU effects are here, mobility was linked to education attainment as transfers are linked to mortality, and membership is defined by time, with different patterns for each individual. There are also differences between the two. Transfers are to some extent dictated by the NNUs and networks, which is not the case for pupil mobility (with the rare exception of expulsions). Infants are transferred for more or less intense care at different times dependent on need, so the patient flows for an NNU will be linked to the type of care it offers as described in chapter 2. No such patterns have been described for the pupil mobility data: while we might expect pupils to move schools more at particular times (e.g. after major examination years) we might not expect these patterns to differ strongly by school in the way that transfer patterns differ by NNU. Transfer patterns are also related to the outcome of the infants as demonstrated in chapter 3. A similar relationship was modelled in detail for pupil mobility data by Leckie (2009) who looked at associations between moves at different time points and educational attainment. I have elected not to adjust for transfers as they are part of care provided. The methods of Leckie (2009) could prove useful if looking at associations between transfer and clinical outcome, although in the educational case moves can only occur at a few time points based on the school year whereas neonatal transfers can occur at any time.

In a provider profiling setting we may only be interested in initial care, in which case

allocation to the first provider may be reasonable, but the MM literature indicates that it is not appropriate if we want to evaluate provider performance regardless of when care occurs. This limitation was not mentioned in any of the studies in the review in section 4.2. Transfers are common for preterm infants, with many infants having multiple transfers so the underestimation of cluster variance when using single membership models may be notable. I will describe methods to investigate this in chapter 5, and to compare the random effects estimated by single and multiple membership models, as existing studies have not done so beyond correlating the ranks.

In section 4.1.4, I discussed how there are particular modelling concerns when using hierarchical models for provider profiling because it is the random effects themselves that drive our inferences about specific providers. Similar concerns exist when choosing the weighting as this may influence the random effects. We may wish all providers to have equal accountability for the patient, or for those providing more care to have a greater share, as is reflected by the standard MM weights which would assign patients to providers based on length of stay. Alternatively, we may want provider effects to reflect the patients under their care at the most crucial points in stay. The suggestion to weight critical times more heavily was made in the context of AMI, stroke and hip fracture patients by Kristoffersen et al. (2012). Other options of fixed-value or variance-based weights described in existing studies cannot be applied to the neonatal data as the timing of hospital episodes is variable: one infant's first admission might be for one day, another's for one week. Whilst several studies found conclusions to be robust to different weights, this may not be the case for the neonatal data because of the nature of transfer patterns, how they vary across NNUs, and their dependence on the health of the infants. When examining sensitivity of cluster effects to different weights, all studies (including the simulation study by Wolff-Smith and Beretvas (2014)) were limited to looking at cluster variance and correlation of the ranks, rather than the effects themselves which is more of interest in this thesis. Therefore in this thesis I aim to implement MM models with a range of weights suitable for neonatal transfers with a focus on estimating NNU effects.

In the next chapter I will describe the methods for the thesis, building on the review of existing methods in this chapter. I will use Bayesian methods in this thesis as, in addition to the advantages for estimating hierarchical models and resulting provider effects described in 4.1, they accommodate the non-nested structure of MM models. A brief background to Bayesian methods will be given in section 5.1. In section 5.2 I will develop a hierarchical model to estimate NNU effects on mortality of the subset of preterm infants who were not transferred, adjusting for the risk factors determined in chapter 3. Based on the findings of this chapter I will explore two and three level models to incorporate variation at the NNU and network level respectively, allowing the mean and variance of NNU effects to vary by NNU level, and using different assumptions for

the random effects distribution. Next I will use MM models to allow the inclusion of transferred infants. This chapter indicated that choice of weights may be important but that the true weight is unknown. In section 5.3 I will describe a simulation study to investigate how different weights suitable for modelling neonatal transfers might affect individual provider effects, in contrast to existing studies. I will then apply MM models to the neonatal data, incorporating data on the intensity of care as suggested by the findings of this chapter.

Chapter 5

Methods for the thesis

In this chapter I describe the analysis methods used in this thesis, building on the methods reviewed in the previous chapter and incorporating the risk factors selected in chapter 3. This chapter begins with a preliminary introduction to the Bayesian techniques (section 5.1) before moving to the main methods for the thesis. Section 5.2 describes the hierarchical models incorporating neonatal unit (NNU) and network effects but only for non-transferred infants so infants can be assigned to a single NNU. I then set out the methods for the multiple membership models, first applied to explore sensitivity to different weights using simulated data in section 5.3 and then to neonatal data (including transferred infants) in section 5.4.

5.1 Bayesian hierarchical models

In this section I give a brief introduction to the Bayesian view of probability and model estimation, to give both a theoretical and practical basis for the methods used in this thesis. More detail about the standard techniques discussed here can be found in Gelman et al. (2004). Bayesian methods provide a different framework for statistical inference from the framework provided by frequentist methods. In a frequentist setting we assume the parameter θ we are trying to estimate using observed data y takes an unknown true value. This is fixed so we cannot make probability statements about it such as "there is a 95% probability that θ is greater than 0"; it either is or it is not. The probability statements we can make under the frequentist paradigm refer to the experiment we are conducting to estimate the value, not value itself. For example we can calculate the p-value as probability of observing data y (or something more extreme) under some assumed null hypothesis for θ. Correspondingly, we can construct a confidence interval for θ using the data y such that the probability that the interval contains θ is over a

certain percentage. As these are probabilities about what we observe, we can think of them as long-run frequencies were we to repeat our experiment many times.

In a Bayesian setting our uncertainty can be expressed directly as a probability in terms of θ, so we are allowed to make statements such as "there is 95% probability that θ is greater than 0". Our uncertainty about the parameters after the experiment is quantified by the posterior distribution (the probability of the parameter θ given the data y), which is calculated using Bayes' theorem (Bayes and Price, 1763):

$$p(\theta|y) = p(y|\theta)p(\theta)/p(y) \propto p(y|\theta)p(\theta) \tag{5.1}$$

The likelihood, $p(y|\theta)$, is the same as that required by likelihood-based frequentist methods, but additionally we now require a prior, $p(\theta)$. The prior quantifies our beliefs about the distribution of θ prior to seeing the data y. The posterior distribution is what we use to draw inferences. The extent to which the prior influences the posterior depends on the how much information is the prior contributes relative to the data. This makes sense if we consider the probability distributions of the parameters of interest in terms of our belief: if we have some idea about the parameters prior to seeing the data, but the data do not tell us much about the parameters, then the posterior distribution will be similar to the prior. Similarly if we have no idea about the values the parameters might take and we have a reasonable amount of data from our experiment, the posterior will mainly be informed by the data (and as the sample size tends to infinity we arrive at the same numerical estimates as those obtained by maximum likelihood, though the interpretation is different).

The prior distributions are included in the specification of the model. For example in model 4.1 we might assign the priors

$$\begin{aligned}\alpha &\sim N(0, 100000)\\ \beta &\sim N(0, 100000)\end{aligned}$$

For model 4.2 we might assign

$$\begin{aligned}\alpha &\sim N(0, 100000)\\ \beta &\sim N(0, 100000)\\ u_j &\sim N(0, \sigma_u^2)\\ \sigma_u &\sim U(0, 100)\end{aligned}$$

This demonstrates the continuity between single level and hierarchical models in a Bayesian framework which arises because there are no fixed parameters: all parameters

are considered random since they are assigned a distribution. The only difference in the specification is that in the hierarchical model, the distribution of the u_j is governed by a higher level parameter σ_u, known as a *hyperparameter*, which is also assigned a prior distribution (a *hyperprior*).

As mentioned previously, it is reasonable to assume that the provider effects are not identical, nor independent, but similar. This notion is formalised in Bayesian statistics as *exchangeability*. A set of parameters $\theta_1, ..., \theta_n$ are defined as finitely exchangeable if the joint distribution $p(\theta_1, ..., \theta_n)$ remains unchanged for any permutation of the labels $1, ..., n$. That is, if I know the set of providers $1, ..., n$, knowing which provider is number 1, which is number 2 and so forth, does not change my beliefs about $\theta_1, ..., \theta_n$. We do not require independence, so knowledge of some of the parameters can give us information about the others. It is this borrowing of information that results the provider effects being shrunk towards the mean, with more shrinkage where there is less information. Exchangeability is also a crucial part of Bayesian estimation itself as it allows the joint prior for all parameters to be decomposed into a product of conditional distributions, further details can be found in Gelman et al. (2004).

5.1.1 Choice of priors

It is recommended to check the sensitivity to the choice of prior by implementing a *community of priors* (Spiegelhalter et al., 1994) which reflect different positions of belief about the parameters. In practice we do not always have much information on which to base the priors, so they can be set to convey minimal information (e.g. a flat distribution over all possible values). These are usually called *vague* or *reference* priors. For some parameters we may be interested in a transformation, in which case to be truly non-informative a prior has to be so under transformation. Priors can also be *weakly informative*, a term applied to those which are restricted to plausible values but do not express any particular prior belief (Gelman, 2006).

For location parameters such as regression coefficients, a Normal distribution with zero mean and a very large variance is a standard option (Lunn et al., 2012). Such parameters are usually insensitive to choice of vague prior when there is a reasonable amount of data available (Gelman, 2006; Lambert et al., 2005). Scale parameters can be more sensitive, in particular the variance parameters in hierarchical models. Gelman (2006) recommends a uniform prior on the standard deviation with a suitably large upper limit if a proper prior is required (as is the case for the OpenBUGS software used here), provided the number of groups not less than five.

5.1.2 Estimation of Bayesian models

A closed-form of the posterior distribution is only available in a limited number of cases. Beyond this, there are various ways of evaluating the required integrals but the most widely used is Markov Chain Monte Carlo (MCMC). Monte Carlo integration is used to approximate any summary of a probability distribution by calculating that summary across repeated independent draws from the distribution. The accuracy of the approximation increases as the number of samples increases. A Markov Chain is a sequence of random variables such that given the present, the future is independent of the past, so all information about the history of the process is contained within the present state. The key property of interest to Bayesian inference is that Markov chains with particular properties converge to a stationary distribution (Gilks et al., 1996). Say we have our parameter vector of interest θ and a set of initial values $\theta^{(0)}$. If we can construct a transition distribution (the set of probabilities governing movement from one set of samples to the next) for which gives the distribution of $\theta^{(t+1)}$ given $\theta^{(t)}$ such that the resulting Markov chain converges on the posterior distribution $p(\theta|y)$, we will obtain a series of samples from a distribution sufficiently close to the target posterior distribution which are independent of the starting values. The starting values can be chosen using the data (as they are not part of our prior belief) but should be suitably different for each Markov chain to allow convergence to be checked (Lunn et al., 2012). This brings us to the required samples we need to carry out Monte Carlo integration as previously described. Gibbs sampling is one way of proceeding from one set of samples to the next using conditional distributions (Gelfand and Smith, 1990; Geman and Geman, 1984) and is the method implemented in the OpenBUGS software (OpenBUGS Project Management Group, 2014) used in this thesis. Practically speaking, if we have a set of parameters it means we sample the next value for one of these parameters (or a subset of several) at a time conditional on the current value of all the other parameters. This approach can be applied to transformations of parameters, allowing us to obtain measures of uncertainty for quantities whose distribution cannot be described parametrically.

5.1.3 Model convergence and estimation of the joint posterior distribution

We take successive samples until we believe that the Markov chain has converged to a stationary distribution. There are no tests which prove convergence but there are a number of statistics to help assess the likelihood of convergence. It is recommended to use multiple ways of assessing convergence as different methods (Lunn et al., 2012); I will describe a number of methods here. Multiple chains can be run from different

starting points and the sampled values plotted for each parameter over time (*trace plots*). If the paths of the chains are indiscernible from one another and look like a random trace then this is consistent with convergence. The Brooks Gelman Rubin (BGR) statistic compares the posterior variability pooling samples from both chains with the average of the posterior variability measured in each chain separately; this ratio should tend to 1 as convergence is reached (Brooks and Gelman, 1998; Gelman and Rubin, 1992). The Geweke statistic (Geweke, 1991) assumes that the chain has converged by the half way point, and compares the last 50% of the chain with the first 10% (these are typical values used). If these sections of the chain are different then this is taken as evidence that the chain has not converged by 10% of the way through. The Heidelberger-Welch stationarity diagnostic (Heidelberger and Welch, 1983) tests a null hypothesis of convergence, first using the whole chain, then discarding an increasing percentage of the initial part of the chain up to 50%.

Once we believe convergence has been reached we can then discard the MCMC samples up to this point (called *burn in*) and run the MCMC until we have enough samples with which to summarise the posterior distribution with sufficient precision. Even once the MCMC simulation has run long enough to reach stationarity, we will still have correlation from one iteration to the next because of the nature of Markov Chains. Independence is not an assumption of Monte Carlo integration, so our inferences are still valid, but they are less efficient (Gilks et al., 1996). If there is a high level of correlation from one sample to the next, it is more difficult for the sampler to explore the parameter space. Autcorrelation can be reduced by only keeping some of the samples (known as *thinning*), though using the whole chain results in more accurate posterior inferences (Lunn et al., 2012).

5.1.4 Model checking and model comparison

Posterior predictive checks are a way of checking whether predictions produced by the model (using the parameters estimated by the posterior distribution) give results that are similar to the original data (Gelman et al., 1996). If the model is correct we would expect that the observed data would be likely to occur under the posterior distribution. Note that, as with most checks of model fit, the converse does not hold, so the lack of apparent problems does not mean that the model is true. Rather, posterior predictive checks are used to investigate ways in which the model might be deficient. We can make predictions for the same individuals and new clusters, or for new individuals and clusters (or combinations thereof) depending on what we are interested in. To conduct posterior predictive checks we simulate data based on sets of parameters sampled from the joint posterior distribution (which is straightforward when we already have such samples from MCMC). We then choose test statistics with which to compare the real

and replicated data. A range of test statics should be chosen to reflect the various ways in which the model may be inadequate, with a focus on the inferences we are interested in (Gelman et al., 2000).

Quantile-quantile (Q-Q) plots can be used to check the assumed distribution of the random effects.These plot the sample (posterior means of the random effects) against the theoretical quantiles of the desired distribution. Even if the model is true and the underlying provider effects follow the specified distribution, the distribution of posterior means will be underdispersed relative to the estimated random effects variance, as the estimates are shrunk towards the overall mean. It is for this reason that the posterior means themselves do not provide a good estimate of the overall variability (Shen and Louis, 1998).

There are a number of statistics which describe how well the model fits the data. The deviance, $D(\theta)$ is a measure of how likely the data are under the parameters θ and is given by:

$$D(\theta) = -2 \log p(y|\theta)$$

The better the fit of the model, the lower the deviance. This measure does not take into account the complexity of the models; usually we want to quantify the trade-off between improvement in fit and added complexity from extra parameters. The Akaike Information Criterion (AIC) which was used in section 3.4.1 does this by adding twice the number of parameters to the estimated deviance. For Bayesian models the number of model parameters is not clear because the parameters are constrained by the priors (Lunn et al., 2012). We consider instead the *effective number of parameters*, p_D which has a number of alternative definitions (Gelman et al., 2004) but I will use that proposed by Speigelhalter (Spiegelhalter et al., 2002):

$$p_D = \bar{D} - D(\bar{\theta})$$

where $\bar{D}$ is the mean of the posterior distribution of the deviance and $D(\bar{\theta})$ is the deviance of the posterior means of the parameters θ. The Deviance Information Criterion (DIC) is an analagous to the AIC and is defined as:

$$\text{DIC} = \bar{D} + p_D = D(\bar{\theta}) + 2p_D$$

I presented DIC, $\bar{D}$ and p_D for all models.

5.2 Hierarchical analysis of non-transferred infants

5.2.1 Aim

In chapter 4 I described how Bayesian hierarchical models can be used to estimate provider effects where individuals attend a single provider. The aim of this section is to describe the application of these models to estimate NNU effects[1] for mortality of the subset of preterm infants who were not transferred. The purpose of the analysis is to investigate some of the other modelling issues raised by the research question before adding the complexity of transfers. In particular I sought to compare differences between modelling clustering by NNU only, and by NNU and neonatal network and ways of modelling differences across NNU levels. I also aimed to assess sensitivity to choice of prior distributions.

5.2.2 Data

The data used in this chapter are a subset of those described in 2.3.1 restricted to singleton infants who were not transferred postnatally i.e. they were treated in a single NNU. As discussed in chapter 2, all infants with missing data were excluded (the limitations of this will be discussed in chapter 9).

5.2.3 Statistical models

In chapter 3 I selected the infant risk factors for inclusion in the model as gestational age at birth, birth weight, sex, and whether the mother received antenatal steroids. Gestational age and birth weight have interacting, non-linear effects. This was modelled with a 3 knot spline for gestation and a quadratic term for birth weight. Interactions between linear gestation with all birth weight terms, and linear birth weight with all gestation terms were included. The same infant risk factors and functional form were applied in all the models described here.

Level of neonatal unit (special care baby unit (SCBU), local neonatal unit (LNU) and neonatal intensive care unit (NICU) in order of increasing level of care provided, see section 1.2 for definitions) is the only provider-level variable I included. The rationale for this was discussed in chapter 3. As the categorisation of NNU into levels is a key part of the structure of neonatal services and therefore of the research question, I compared

[1] As mentioned in section 4.1.4, the word "effect" is from the random effects terminology. NNU effects are not wholly caused by the NNU, as they may be attributable to unmodelled factors which are not under the control of the NNU.

models with and without NNU level as a covariate, and with and without complex variation by NNU level. In chapter 4 I described how hierarchical models can be used to describe clustering at more than two levels. I therefore compared models with both two levels (infants within NNUs) and three levels (infants within NNUs within networks).

Similarly to the notation for the risk model in section 3.4 and the hierarchical models introduced in section 4.1, let p_{ij} denote the probability that infant i in NNU j dies before discharge, with $i = 1, ..., N$. Let the gestational age spline terms for infant i be denoted by $g_{ij}^{(1)}$ and $g_{ij}^{(2)}$, birth weight by b_{ij}, with the variables mean centred and standardised as before. Binary indicators for male sex and whether the mother received any antenatal steroids are given by m_{ij} and s_{ij} respectively. As before the NNU effects are given by u_j and the between-NNU variance by σ_u^2. To incorporate the level of NNU into the model specification we use the notation level(j) to mean the level of NNU j.

The two-level models are described as follows.

Model 1 A two-level model (infants within NNUs), adjusted for infant risk factors only:

$$\begin{aligned} logit(p_{ij}) &= \beta_1 g_{ij}^{(1)} + \beta_2 g_{ij}^{(2)} + \beta_3 b_{ij} + \beta_4 b_{ij}^2 \\ &\quad + \beta_5 g_{ij}^{(1)} b_{ij} + \beta_6 g_{ij}^{(1)} b_{ij}^2 + \beta_7 b_{ij} g_{ij}^{(2)} \\ &\quad + \beta_8 m_{ij} + \beta_9 s_{ij} + \mu_j && i = 1, ..., N \\ \mu_j &= \alpha + u_j && j = 1, ..., J \end{aligned} \tag{5.2}$$

Priors

$$\begin{aligned} \alpha &\sim N(0, 100000) \\ \beta_s &\sim N(0, 100000) && s = 1, ..., 9 \\ u_j &\sim N(0, \sigma_u^2) \end{aligned}$$

Hyperprior

$$\sigma_u \sim U(0, 5)$$

Model 2 A two-level model (infants within NNUs), adjusted for infant risk factors and NNU level, taking LNU as the baseline category:

$$\begin{aligned} logit(p_{ij}) &= \beta_1 g_{ij}^{(1)} + \beta_2 g_{ij}^{(2)} + \beta_3 b_{ij} + \beta_4 b_{ij}^2 \\ &\quad + \beta_5 g_{ij}^{(1)} b_{ij} + \beta_6 g_{ij}^{(1)} b_{ij}^2 + \beta_7 b_{ij} g_{ij}^{(2)} \\ &\quad + \beta_8 m_{ij} + \beta_9 s_{ij} + \mu_j && i = 1, ..., N \\ \mu_j &= \alpha + \gamma_1 \mathbb{1}_{\text{SCBU}}(j) + \gamma_2 \mathbb{1}_{\text{NICU}}(j) + u_j && j = 1, ..., J \end{aligned}$$

Priors

$$\begin{aligned}
\alpha &\sim N(0, 100000) \\
\beta_s &\sim N(0, 100000) && s = 1, ..., 9 \\
\gamma_t &\sim N(0, 100000) && t = 1, 2 \\
u_j &\sim N(0, \sigma_u^2)
\end{aligned}$$

Hyperpriors

$$\sigma_u \sim U(0, 5)$$

where:

$$\mathbb{1}_{\text{type}}(j) = \begin{cases} 1 & \text{if level}(j) = \text{type} \\ 0 & \text{if level}(j) \neq \text{type} \end{cases} \quad \text{type} \in \{\text{SCBU, NICU}\}$$

Model 3 A two-level model (infants within NNUs), adjusted for infant risk factors and NNU level, with level-specific variance for the NNU effects:

$$\begin{aligned}
logit(p_{ij}) &= \beta_1 g_{ij}^{(1)} + \beta_2 g_{ij}^{(2)} + \beta_3 b_{ij} + \beta_4 b_{ij}^2 \\
&\quad + \beta_5 g_{ij}^{(1)} b_{ij} + \beta_6 g_{ij}^{(1)} b_{ij}^2 + \beta_7 b_{ij} g_{ij}^{(2)} \\
&\quad + \beta_8 m_{ij} + \beta_9 s_{ij} + \mu_j && i = 1, ..., N \\
\mu_j &= \alpha + \gamma_1 \mathbb{1}_{\text{SCBU}}(j) + \gamma_2 \mathbb{1}_{\text{NICU}}(j) + u_j && j = 1, ..., J
\end{aligned}$$

Priors

$$\begin{aligned}
\alpha &\sim N(0, 100000) \\
\beta_s &\sim N(0, 100000) && s = 1, ..., 9 \\
\gamma_t &\sim N(0, 100000) && t = 1, 2 \\
u_j &\sim N(0, \sigma_{u^{(\text{level}(j))}}^2)
\end{aligned}$$

Hyperprior

$$\sigma_{u^{(\text{level}(j))}} \sim U(0, 5) \qquad \text{level}(j) \in \{\text{SCBU, LNU, NICU}\}$$

Similarly to equation 4.3, for the three level model we use the ijk subscript to denote infant i in NNU j in network k, where $k = 1, ..., K$.

Model 4 A three-level model (infants within NNUs within networks), adjusted for infant risk factors only:

$$
\begin{aligned}
logit(p_{ijk}) &= \beta_1 g_{ijk}^{(1)} + \beta_2 g_{ijk}^{(2)} + \beta_3 b_{ijk} + \beta_4 b_{ijk}^2 \\
&\quad + \beta_5 g_{ijk}^{(1)} b_{ijk} + \beta_6 g_{ijk}^{(1)} b_{ijk}^2 + \beta_7 b_{ijk} g_{ijk}^{(2)} \\
&\quad + \beta_8 m_{ijk} + \beta_9 s_{ij} + \mu_j && i = 1, ..., N \\
\mu_{jk} &= \nu_k + u_{jk} && j = 1, ..., J \\
\nu_k &= \alpha + v_k && k = 1, ..., K
\end{aligned}
$$

Priors

$$
\begin{aligned}
\alpha &\sim N(0, 100000) \\
\beta_s &\sim N(0, 100000) && s = 1, ..., 9 \\
u_{jk} &\sim N(0, \sigma_u^2) \\
v_k &\sim N(0, \sigma_v^2)
\end{aligned}
$$

Hyperpriors

$$
\begin{aligned}
\sigma_u &\sim U(0, 5) \\
\sigma_v &\sim U(0, 5)
\end{aligned}
$$

Model 5 A three-level model (infants within NNUs within networks), adjusted for infant risk factors and NNU level:

$$
\begin{aligned}
logit(p_{ijk}) &= \beta_1 g_{ijk}^{(1)} + \beta_2 g_{ijk}^{(2)} + \beta_3 b_{ijk} + \beta_4 b_{ijk}^2 \\
&\quad + \beta_5 g_{ijk}^{(1)} b_{ijk} + \beta_6 g_{ijk}^{(1)} b_{ijk}^2 + \beta_7 b_{ijk} g_{ijk}^{(2)} \\
&\quad + \beta_8 m_{ijk} + \beta_9 s_{ij} + \mu_j && i = 1, ..., N \\
\mu_{jk} &= \nu_k + \gamma_1 \mathbb{1}_{\text{SCBU}}(j) + \gamma_2 \mathbb{1}_{\text{NICU}}(j) + u_{jk} && j = 1, ..., J \\
\nu_k &= \alpha + v_k && k = 1, ..., K
\end{aligned}
$$

Priors

$$
\begin{aligned}
\alpha &\sim N(0, 100000) \\
\beta_s &\sim N(0, 100000) && s = 1, ..., 9 \\
\gamma_t &\sim N(0, 100000) && t = 1, 2 \\
u_{jk} &\sim N(0, \sigma_u^2) \\
v_k &\sim N(0, \sigma_v^2)
\end{aligned}
$$

Hyperpriors

$$
\begin{aligned}
\sigma_u &\sim U(0, 5) \\
\sigma_v &\sim U(0, 5)
\end{aligned}
$$

Model 6 A three-level model (infants within NNUs within networks), adjusted for infant risk factors and NNU level, with level-specific variance for the NNU effects:

$$\begin{aligned}
logit(p_{ijk}) &= \beta_1 g^{(1)}_{ijk} + \beta_2 g^{(2)}_{ijk} + \beta_3 b_{ijk} + \beta_4 b^2_{ijk} \\
&\quad + \beta_5 g^{(1)}_{ijk} b_{ijk} + \beta_6 g^{(1)}_{ijk} b^2_{ijk} + \beta_7 b_{ijk} g^{(2)}_{ijk} \\
&\quad + \beta_8 m_{ijk} + \beta_9 s_{ij} + \mu_j && i = 1, ..., N \\
\mu_{jk} &= \nu_k + \gamma_1 \mathbb{1}_{\text{SCBU}}(j) + \gamma_2 \mathbb{1}_{\text{NICU}}(j) + u_{jk} && j = 1, ..., J \\
\nu_k &= \alpha + v_k && k = 1, ..., K
\end{aligned}$$

Priors

$$\begin{aligned}
\alpha &\sim N(0, 100000) \\
\beta_s &\sim N(0, 100000) && s = 1, ..., 9 \\
\gamma_t &\sim N(0, 100000) && t = 1, 2 \\
u_{jk} &\sim N(0, \sigma^2_{u^{(\text{level}(j))}}) \\
v_k &\sim N(0, \sigma^2_v)
\end{aligned}$$

Hyperpriors

$$\begin{aligned}
\sigma_{u^{(\text{level}(j))}} &\sim U(0, 5) && \text{level}(j) \in \{\text{SCBU, LNU, NICU}\} \\
\sigma_v &\sim U(0, 5)
\end{aligned}$$

OpenBUGS code for these models is provided in appendix D.

5.2.4 Prior distributions

In the models above I have specified a Normal distribution for the NNU and network effects and a Uniform (0,5) prior for standard deviation of the random effect distributions. The upper limit for the Uniform distribution is equivalent to an upper limit of 25 for the variance of the log odds ratios. This corresponds to a 95% range of 0.007 to 148 on the odds ratio scale, so still much wider than we would expect and so should be sufficiently far from the tails of the posterior distribution. As provider effects can be sensitive to the choice of priors and hyperpriors I conducted several sensitivity analyses based on the findings of chapter 4. I specified a t distribution with 4 degrees of freedom for the NNU effects (in place of the Normal distribution specified above). I did not use an alternative distribution for the network effects as there are too few networks to compare fit. Regarding hyperpriors for the standard deviation of NNU and network effects, I replaced the $U(0, 5)$ in the main analysis with a Uniform on a wider range of $U(0, 5)$. For the fixed parameters (intercept and regression coefficients) I have used a

vague Normal prior with a large variance.

5.2.5 Model estimation and checking

All analyses were run in R version 3.2.3 (R Core Team, 2015) OpenBUGS version 3.2.3 (OpenBUGS Project Management Group, 2014). I ran two MCMC chains with dispersed initial values and used trace plots and the BGR statistic (Brooks and Gelman, 1998; Gelman and Rubin, 1992) to determine the initial burn-in period. I ran the MCMC chains until the upper limit of the confidence interval for the BGR was less than 1.1 for all regression coefficients, random effect variances and a selection of NNU effects. I then doubled the number of iterations and used the Geweke (Geweke, 1991) and Heidelberger-Welch (Heidelberger and Welch, 1983) tests to assess whether any more initial iterations should be discarded as burn in. I conducted both tests on both MCMC chains and discarded further iterations if there was evidence of non-convergence for any single parameter from more than one source. The rationale for this approach was that these procedures both involve hypothesis tests on multiple parameters so, at a significance level of 0.05, we would expect one in 20 tests to be statistically significant under convergence. I then continued running the MCMC iterations until the Monte Carlo standard error (a measure of the accuracy with which the samples estimate the posterior) had been reduced to 5% of the posterior standard deviation (SD) for all parameters. This is equivalent to having 400 independent samples (called the *effective sample size* (ESS)) which is satisfactory for most purposes (Lunn et al., 2012). For this thesis I am not reliant on tail probabilities, for which a much larger ESS is required for the necessary precision, so the standard approach is adequate. All model checks were done using the CODA package for R (Plummer et al., 2006).

To conduct posterior predictive checks I drew 1000 replications of the data, predicting for the same infants within the same NNUs (i.e. simulating $Y_{ijk}^{new} = \text{Bernoulli}(p_{ijk})$ for each infant i in NNU j in network k in the dataset, with p_{ij} as defined for each model 1 to 6). I compared the mortality rates at NNUs, neonatal networks and by NNU level for the observed and replicated data as these quantities relate to how the structure of neonatal services is incorporated in the model. I used graphical checks to detect discrepancies, plotting histograms of the predicted mortality rate across simulated datasets along with the observed rate as a vertical line.

Quantile-quantile (Q-Q) plots were used to check the assumed distribution of the NNU effects and the deviance, effective parameters and DIC compared across models.

5.2.6 Presenting results

Infant and provider characteristics were presented as mean posterior odds ratios and 95% credible intervals (the 2.5th to 97.5th percentiles of the posterior distribution). Density plots were used to illustrate the posterior distribution for the variance parameters. The mean of the standard deviations of the posterior distributions of individual NNU effects (averaged across NNUs) was presented to give an estimate of the uncertainty in estimating provider effects. The variance partitioning coefficient (VPC) was used to estimate the proportion of total variation attributable to differences between providers. For continous variables partitioning the variance is straightforward, but for binary variables the link function (in this case the logit) means amount of variation is expressed on different scales at different levels of the model (as a probability at the individual level and continuously at the cluster level). Here I used the latent variable approach (also called the linear threshold method) to calculate the VPC (Goldstein et al., 2002; Snijders and Bosker, 1999). We consider the observed binary outcome as resulting from a latent continuous variable such that the observed value is 0 if the latent variable is below a threshold, and 1 if it is above the threshold. The latent variable (under the logistic model used here) follows a logistic distribution with variance $\pi^2/3$, which we then take to be the individual-level variance on the continuous scale of the latent variable. For the two-level models, the VPC for the between NNU variance σ_u^2 is therefore calculated as

$$VPC_u = \frac{\sigma_u^2}{\sigma_u^2 + \frac{\pi^2}{3}}$$

For the three-level models the VPC for the between network variation takes a similar form to the above but with the variation at three levels in the denominator:

$$VPC_v = \frac{\sigma_v^2}{\sigma_u^2 + \sigma_v^2 + \frac{\pi^2}{3}}$$

In the three level model we can also consider the VPC for NNUs and networks combined:

$$VPC_{u_{diff}} = \frac{\sigma_u^2 + \sigma_v^2}{\sigma_u^2 + \sigma_v^2 + \frac{\pi^2}{3}}$$

The posterior mean NNU effects were compared using pair-wise scatter plots and correlations. Density plots of the posterior means by level were presented for each model. Finally plots of the smallest and largest estimates of u_j across models for each NNU were plotted, with 95% credible intervals, by NNU level to illustrate how results for individual NNUs vary across models. Neonatal units and networks are not named as this was an exploratory analysis to investigate methodology.

5.3 Multiple membership analysis using simulated data

5.3.1 Aim

In chapter 4, I reviewed how transferred patients have been handled using hierarchical models in the provider profiling literature, and introduced multiple membership (MM) models as a possible alternative. MM models require the specification of weights to describe how individuals are assigned to each cluster. As described in section 4.3.2, the weighting scheme used to analyse the data affects the estimates of σ_u^2, u_j and the VPC because the assumed cluster-level variance contribution for an individual will vary by scheme. For provider profiling purposes the weights chosen should reflect the inferences we want to make about the provider effects. However we need to acknowledge that these weights may attribute variation in a different way to the underlying processes that gives rise to the observed data. The aim of this section is to use simulated data to investigate the differences in variance parameters and provider effects under a variety of data scenarios and analysis models.

5.3.2 Weighting schemes

A standard option for the weights is the proportion to time spent in each cluster (length of stay for hospital patients), as described in section 4.3. I discussed there that, for provider profiling, we may also want to consider weights that reflect the intensity of care, which could vary throughout the patient journey. Although infants may not be in the first neonatal unit for long, initial care prior to transfer is considered crucial. Indeed the term *golden hour* has been borrowed from adult emergency medicine to refer to the first postnatal hour (Annibale and Bissinger, 2010). While, as with adult trauma, the evidence to support the importance of this exact hour is limited, several studies suggests that differences in initial care lead to variations in mortality (Empana et al., 2003; Marlow et al., 2014; Watson et al., 2014). Infants have to be resuscitated and stabilised appropriately and a transfer organised(Fenton et al., 2004). More intensive care is performed earlier during stay (section 2.3.2),and the risk of mortality is higher (section 2.3.4). Therefore there is a clinical argument for placing greater weight on the earlier part of stay in the neonatal setting so this led the weighting choice for the simulation study.

Since the sum of the weights over the whole stay is one, and since the total length of stay varies, I specified the weight as the area under the Beta probability density function (pdf) as a function of the proportion of total stay. The Beta distribution has two parameters which describe the shape, here denoted by a and b. For an episode with

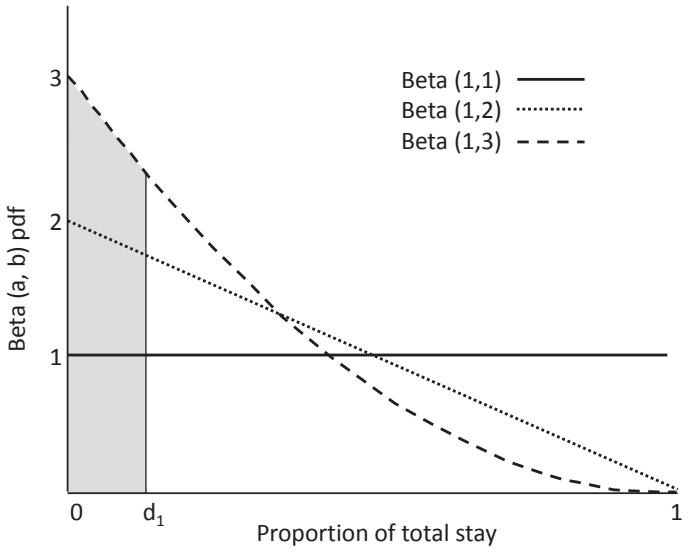


Figure 5.1: Multiple membership weights described by Beta distributions. The weights are given by the area under the line for that part of stay. For example, for the first episode ending at time d_1(shaded area), the Beta(1,3) weighting gives the greatest weight to the hospital and the Beta(1,1) weighting the smallest weight.

admission time t_1 and discharge time t_2 (with these times expressed as a proportion of total stay) where $f_B(t; a, b)$ and $f_B(t; a, b)$ denote the pdf and cumulative distribution function (cdf) of the Beta distribution with parameters a and b, then the weight w for this episode is

$$w = \int_{t_1}^{t_2} f_B(t; a, b)\, dt = F_B(t_2; a, b) - F_B(t_1; a, b) \tag{5.3}$$

Figure 5.1 shows the pdf for Beta(1,1), Beta(1,2) and Beta(1,3) distributions. The Beta(1,1) is flat, so the weighting formula above is equivalent to the standard option of apportioning weight according to length of stay. A Beta(1,2) distribution gives more weight to the earlier part of stay, equivalent to half the weight allocated to the first 30% of stay. As the second Beta parameter increases, the weight becomes more unevenly distributed, with the extreme scenario being all weight allocated to the first provider. A Beta(1,3) distribution is equivalent to half the weight allocated to the first 20% of stay.

To show the difference in cluster-level variance contribution across weighting schemes, suppose we have an individual who spends an equal amount of time in two clusters. Under the Beta(1,1) weight their variance contribution will be $0.5^2\sigma_u^2 + 0.5^2\sigma_u^2 = 0.5\sigma_u^2$, under Beta(1,2) $0.75^2\sigma_u^2 + 0.25^2\sigma_u^2 = 0.625\sigma_u^2$ and under Beta(1,3) $0.875^2\sigma_u^2 + 0.125^2\sigma_u^2 = 0.781\sigma_u^2$.

5.3.3 Data

A simple model was followed with only two levels (individuals within provider) and no individual-level or provider-level covariates. I simulated data from 15,000 infants across 100 providers.

All infants were randomly assigned to a first provider from the 100 providers with equal probability. 25% of the infants were randomly selected to be transferred once. The second provider was randomly selected from all providers other than the first. Transfer times as a proportion of total stay were selected using a Uniform (0,1) distribution. This is a simpler transfer than that observed in the neonatal data as described in chapter 2; limitations of this will be discussed in section 7.5.

The log odds of mortality for providers were drawn from a Normal distribution with a mean of -2.2 and standard deviation 0.4, corresponding to 95% of underlying provider-specific mortality rates lying between 5% and 20%. Therefore the log odds ratios representing provider effects, denoted u_j were drawn from a Normal distribution with mean 0 and standard deviation 0.4. The outcome for infant i was generated as a Bernoulli random variable with probability p_i given by

$$\begin{aligned} logit(p_i) &= -2.2 + u_j && \text{for individual } i \text{ treated in a single provider } j \\ logit(p_i) &= -2.2 + w_{i1}u_{j1} + w_{i2}u_{j2} && \text{for individual } i \text{ treated in two providers } j_1 \text{ and } j_2 \end{aligned}$$

The weights w_{i1} and w_{i2} were determined by a Beta(a, b) distribution as described in equation 5.3. For w_{i1}, $t_1 = 0$ and t_2 is the time of transfer for individual i. For w_{i2}, t_1 is the time of transfer for individual i and $t_2 = 1$.

For simulating the data we assumed three scenarios: weight proportional to length of stay (LOS) which is equivalent to Beta(1,1); moderate emphasis on the early part of stay calculated with a Beta(1,2) distribution; and a strong emphasis on the early part of stay using a Beta(1,3) distribution. For convenience I will label these *LOS*, *Beta2* and *Beta3* respectively. These three options were chosen because they seemed the most plausible for the neonatal data compared with alternatives allocating greater weight to the middle or later parts of stay. Two "early" options were chosen rather than one because it allows us to see if results are sensitive to lesser misspecification. For each of the three scenarios I simulated 200 datasets (600 datasets in total). The datasets were identical across the four analysis models, but vary across the three weighting scenarios as well as within scenarios. The stochastic elements were the Bernoulli survival indicator, the transfer times and the allocation of individuals to providers. Provider effects were generated once only as these were examined across models and scenarios.

5.3.4 Statistical models and estimation

For analysing the data, in addition to the three weighting schemes used to generate the data, an additional analysis attributing each individual to the first provider only was carried out to investigate the impact of ignoring the multiple membership. This weighting is labelled *First* for convenience. This gave twelve sets of results in total: three scenarios (*LOS*, *Beta2*, *Beta3*) each analysed with four different weightings (*LOS*, *Beta2*, *Beta3*,*First*). Adopting the more general notation for the weights that was used in 4.3 and hierarchical centring for consistency with analysis models in section 5.4, we let $U(i)$ denote the set of providers treating individual i. As in section 5.2, Bayesian methods were used to estimate the models, with the same vague priors applied for the intercept and random effect variance. Vague priors were used to mimic the lack of knowledge about provider effects for real data analysis. The analysis models are given by:

$$\begin{aligned} logit(p_i) &= \sum\nolimits_{j \in \mathrm{U}(i)} w_{ij}\mu_j & i &= 1, ..., N \\ \mu_j &= \alpha + u_j & j &= 1, ..., J \end{aligned}$$

Prior

$$\begin{aligned} \alpha &\sim N(0, 100000) \\ u_j &\sim N(0, \sigma_u^2) \end{aligned}$$

Hyperprior

$$\sigma_u \sim U(0, 5)$$

with weights as described in equation 5.3, setting t_1 as 0 and t_2 as the time of transfer for individual i. For the *First* analysis, all weight is assigned to the first provider. Model estimation was as described in section 5.2.5.

5.3.5 Presenting results

The posterior mean averaged over the M simulated datasets was presented for α and σ_u for each of the 12 analysis-scenario combinations. A variety of measures were used to describe the differences between the estimates produced by the analysis models and the parameters which generated the data. Percentage bias, empirical standard errors (SE), mean square error (MSE) and coverage for α and σ_u are presented, with these measures defined as follows (Burton et al., 2006). If β is the parameter of interest, $\hat{\beta}_m$ is the posterior mean of β from simulated dataset $m, m = 1, ..., M$ and $\bar{\hat{\beta}}$ is the mean of

$\hat{\beta}$ over the M simulations then he percentage bias is given by

$$\frac{\bar{\hat{\beta}} - \beta}{\beta} \times 100\% \tag{5.4}$$

the empirical SE is given by

$$\sqrt{\frac{1}{M-1}\sum_{m=1}^{M}(\hat{\beta}_m - \bar{\hat{\beta}})^2} \tag{5.5}$$

and the mean square error is given by

$$(\bar{\hat{\beta}} - \beta)^2 + SE(\hat{\beta})^2 \tag{5.6}$$

The coverage is percentage of the M datasets where the 95% credible interval includes the known true parameter.

To examine the impact on provider effects we present the bias, MSE and coverage averaged over all 100 providers. The average posterior SD over all providers is presented as a measure of the uncertainty around provider effects. To investigate whether the Deviance Information Criterion (DIC) (Spiegelhalter et al., 2002) discriminates between the different models, I present the mean DIC and the proportion of datasets where the correct analysis model is chosen based on DIC (applying a minimum difference of 5 (Spiegelhalter et al., 2003)).

5.4 Multiple membership analysis using neonatal data

5.4.1 Aim

The aim of this section is to describe methods to estimate NNU effects on mortality for singleton preterm infants, including those who were transferred. I applied MM models described in section 4.3, building on the methods for the hierarchical analysis of non-transferred infants described in section 5.2. I explored sensitivity to different weights, as was done with simulated data in section 5.3.

5.4.2 Weighting schemes

In addition to the four analysis weightings used in the simulation study (*LOS*, *Beta2*, *Beta3*,*First*) I applied two other weightings which are relevant for provider profiling purposes. As explained in chapter 1, for each day an infant is in the neonatal unit their care is classified as *intensive care* (IC), *high dependency care* (HDC) or *special care*

(SC) in decreasing order of intensity. Therefore we can augment the *LOS* weighting described in section 5.3 by assigning weight according to the intensity of care received each day. This weighted length of stay approach is labelled *WLOS* for convenience. For each infant, each day in neonatal care (across all NNUs providing care) was assigned a weight of 1 for IC, 0.5 for HDC and 0.25 for SC. These ratios are the nurse staffing ratios recommended by the British Association of Perinatal Medicine (1:1 nurse to infant ratio for IC, 1:2 for HDC and 1:4 for SC, see section 1.2). Letting d_{ij}^{IC} denote the number of IC days infant i has in NNU j, and d_{ij}^{IC} the total number of IC days for infant i (with similar definitions for HDC and SC) then the weights are then given by:

$$w_{ij} = \frac{d_{ij}^{IC} + 0.5d_{ij}^{HDC} + 0.25d_{ij}^{SC}}{d_{i}^{IC} + 0.5d_{i}^{HDC} + 0.25d_{i}^{SC}}$$

As discussed in chapter 4, in some provider profiling analyses which did not use hierarchical models, individuals were assigned to all providers that treated them. A similar approach for multiple membership models is to assign individuals to all providers equally regardless of length of stay, i.e.

$$w_{ij} = 1/N_i \quad \forall j \in U(i)$$

where $U(i)$ denotes the set of N_i providers treating individual i. This weighting is given the label *Equal*. A summary of all six weights is shown in table 5.1.

Table 5.1: Weightings for multiple membership models for simulated and neonatal data

Label	Description
LOS	The outcome is attributed in proportion to the length of stay
Beta2	The outcome is attributed using the Beta(1,2) weight
Beta3	The outcome is attributed using the Beta(1,3) weight
First	The outcome is attributed to the first provider only
WLOS *(neonatal data only)*	The outcome is attributed according to length of stay weighted by the type of care received, using recommended nurse staffing ratios
Equal *(neonatal data only)*	The outcome is attributed equally to all NNUs caring for the infant, regardless of length of stay or intensity of care

5.4.3 Data

The data used in this chapter are a subset of those described in 2.3.1 restricted to singleton infants with no missing data, as for the analysis of non-transferred infants described in 5.2. I also excluded any infants without a complete episodic record with all

care in English NNUs. Episodic records were determined as incomplete if there was a missing episode of care based on the episode number, or by discrepant details (more than 24 hours between discharge from one NNU and admission to another, or if the discharge destination NNU was different to the NNU of the next episodic record). Infants who had an episode where the episode number, admission or discharge times, the NNU providing care or daily care level data were missing were also excluded. A brief description of this cohort is provided in chapter 7. I summarised the distribution of the total weight for all infants across NNU levels and episodes of care to illustrate how the weighting schemes differ in practice.

5.4.4 Statistical models and estimation

As for the analysis of non-transferred infants, risk factors included in the model were those determined in chapter 3: gestational age at birth, birth weight, sex, and whether the mother received antenatal steroids, with spline terms and interactions as described in 5.2. Modelling strategies for neonatal network effects, variation by level of NNU, and prior distributions for the random effects were guided by the findings of the analysis described in 5.2. Using Model 1 (equation 5.2) as an example we have:

$$\begin{aligned} logit(p_i) &= \beta_1 g_i^{(1)} + \beta_2 g_i^{(2)} + \beta_3 b_i + \beta_4 b_i^2 \\ &\quad + \beta_5 g_i^{(1)} b_i + \beta_6 g_i^{(1)} b_i^2 + \beta_7 b_i g_i^{(2)} \\ &\quad + \beta_8 m_i + \beta_9 s_i + \sum\nolimits_{j \in \mathrm{U}(i)} w_{ij} \mu_j \qquad i = 1, ..., N \\ \mu_j &= \alpha + u_j \qquad j = 1, ..., J \end{aligned}$$

Priors

$$\begin{aligned} \alpha &\sim N(0, 100000) \\ \beta_s &\sim N(0, 100000) \qquad s = 1, ..., 9 \\ u_j &\sim N(0, \sigma_u^2) \end{aligned}$$

Hyperprior

$$\sigma_u \sim U(0, 5)$$

Note that, as described when these models were introduced in 4.3, we subscript any quantities at the infant level with i rather than ij as the infant is no longer nested in a single NNU. Multiple membership for the other models follows in the same way (the μ_j is replaced by $\sum_{j \in \mathrm{U}(i)} w_{ij} \mu_j$).

Six different models were run, one for each of the weighting schemes described in table 5.1. Model estimation and convergence were as described in section 5.2.5. The three-level models allow for the cross-network transfers described in chapter 2 as the

network effects are specified in the μ_j and weighted accordingly (see equation 4.6)

5.4.5 Presenting results

In addition to the results and model fit statistics presented in 5.2.6, I investigated patterns by transfer rate, considering proportion transferred, transfers in and out, and transfers within 24 hours. The rationale for this is that any sensitivity to different weights is most likely to be seen in NNUs with the highest transfer activity, and analysis in chapter 2 showed differences between early and late transfers, with early transfers generally being for a higher level of care.

5.5 Summary

In this chapter I described the statistical methods for the main analyses of the thesis. First I set out a series of models to investigate some of the statistical issues raised by the research question before considering transfers. In particular I sought to compare differences between modelling clustering by NNU only, by NNU and neonatal network, and ways of modelling differences across NNU levels. I also aimed to assess sensitivity to choice of prior distributions. Then I described a simulation study to assess how sensitive parameters of interest are under a variety of data scenarios and analysis models. I compared the usual weight applied in MM models (proportional with time spent in each cluster i.e. length of stay) with weighting more heavily on the earlier part of stay in line with the findings of chapter 2, and with allocation to first provider to reflect standard practice in the profiling literature. Finally I described the application of MM models with a series of different weights to the neonatal data, building on the models developed in the non-transferred infants. In addition to the weights applied in the simulation study, I weighted the length of stay at each NNU according to the category of care provided, and a simple measure assigning infants to each NNU equally. The results of applying the methods of this chapter are presented for the whole population in chapters 6 and 7. A case study for a single NNU is presented in chapter 8 and any methods pertaining solely to that chapter are described there.

Chapter 6

Results I: Non-transferred infants

In this chapter I present the results of applying the methods described in section 5.2. I compare the six Bayesian hierarchical models to estimate neonatal unit (NNU) effects on mortality of non-transferred infants. The purpose of this chapter is to explore how the different ways of modelling the data structure affect the results before including transferred infants in chapter 7. I first recap the methods for the chapter (section 6.1), describe the data (section 6.2) and the results of model checking procedures (section 6.3). Associations between the infant and provider characteristics are presented in section 6.4. The key findings concern the NNU and network effects; these are presented in section 6.5. The overall distribution and standard deviation estimates are compared across models, along with individual NNU and network effects. The interpretation of the results, including implications for the transfers analysis is presented in section 6.6.

6.1 Methods recap

This section is a brief summary of the methods underlying the results for this chapter; full details were given in chapter 5. Six Bayesian hierarchical models were estimated using Markov Chain Monte Carlo methods. The response variable was death before discharge from neonatal care. The population was all singleton infants born below 32 weeks gestation between 2011 and 2013 and admitted to an NNU in England contributing data to the National Neonatal Research Database (NNRD) who were not transferred during the course of neonatal care.

Model 1 A two-level model (infants within NNUs), adjusted for infant risk factors based on the findings of chapter 3 (gestational age (as a restricted cubic spline with 3 knots), birth weight (as linear and quadratic), all interactions between

gestation and birth weight involving linear terms, sex and whether the mother received antenatal steroids)

Model 2 As Model 1, but including level of NNU (special care baby unit (SCBU), local neonatal unit (LNU) or neonatal intensive care unit (NICU), in increasing order of level of care offered) as an NNU-level variable in the model

Model 3 As Model 2, but allowing the between-NNU variance to differ by level of NNU (*complex variation*)

Model 4 As Model 1 but with a third level of the data hierarchy corresponding to neonatal networks

Model 5 As Model 2 but with a third level of the data hierarchy corresponding to neonatal networks

Model 6 As Model 3 but with a third level of the data hierarchy corresponding to neonatal networks

I considered both a Normal distribution and a t distribution with four degrees of freedom for the NNU effects, and a Normal distribution was assumed for the network effects. Vague priors were used for all parameters. As results can be sensitive to choice of prior for the random effects variance, and a U(0,100) prior for the between-NNU standard deviation was used a sensitivity analysis to the main U(0,5) analysis. The focus of interest is comparing results across models is to see whether inferences regarding NNU and network effects change, and to inform modelling choices for chapter 7.

6.2 Data

Table 6.1 shows the infant characteristics of the cohort analysed in this chapter. This cohort is a subset of the cohort described in section 2.3.1 restricted to non-transferred infants only and excluding the 147 non-transferred infants for whom data on antenatal steroids was missing. As these infants were only treated in one NNU, data are presented by the level of NNU to illustrate differences in the population. Higher level NNUs have a higher proportion of lower gestation and lower birth weight infants and higher mortality rates, in line with the description in chapter 1 and the analysis of all infants in chapter 2.

Table 6.1: Description of cohort for analysis of non-transferred infants by NNU level, all data shown as n(%)

		SCBU	LNU	NICU	Total
Gestational age (weeks$^{+\text{days}}$)	$22^{+0} - 23^{+6}$	5(0.9)	20(0.5)	192(3.6)	217(2.1)
	$24^{+0} - 25^{+6}$	5(0.9)	71(1.7)	622(11.7)	698(6.9)
	$26^{+0} - 27^{+6}$	9(1.6)	436(10.3)	930(17.5)	1375(13.6)
	$28^{+0} - 29^{+6}$	94(16.6)	1312(30.9)	1362(25.7)	2768(27.4)
	$30^{+0} - 31^{+6}$	454(80.1)	2407(56.7)	2194(41.4)	5055(50.0)
Birth weight (g)	< 500	0(0.0)	7(0.2)	61(1.2)	68(0.7)
	$500 - 999$	25(4.4)	619(14.6)	1789(33.8)	2433(24.1)
	$1000 - 1499$	235(41.4)	2177(51.3)	2142(40.4)	4554(45.0)
	$1500 - 1999$	282(49.7)	1333(31.4)	1224(23.1)	2839(28.1)
	$\geqslant 2000$	25(4.4)	110(2.6)	84(1.6)	219(2.2)
Sex	Female	253(44.6)	1940(45.7)	2426(45.8)	4619(45.7)
	Male	314(55.4)	2306(54.3)	2874(54.2)	5494(54.3)
Steroids*	No	71(12.5)	530(12.5)	655(12.4)	1256(12.4)
	Yes	496(87.5)	3716(87.5)	4645(87.6)	8857(87.6)
Death before discharge	Survived	539(95.1)	4015(94.6)	4456(84.1)	9010(89.1)
	Died	28(4.9)	231(5.4)	844(15.9)	1103(10.9)

* Any dose of antenatal steroids given to the mother

6.3 Model checking

6.3.1 Convergence

All models were run with a burn-in of 20,000 iterations and a subsequent posterior sample of 150,000 iterations for models 1,2 and 3 and 200,000 iterations for models 4,5 and 6. All models appeared to converge though mixing was poor for some parameters, hence a large number of posterior samples were required to achieve the minimum effective sample size. Model details and convergence assessment were as described in section 5.2; the Monte Carlo standard error and effective sample size are provided in appendix E.

6.3.2 Choice of random effect distribution and priors

Modelling the NNU effects with a t distribution with 4 degrees of freedom gave a better fit than the Normal distribution (see section 6.5), therefore results from the t-4 models are presented as the main results. An alternative prior of U(0,100) on the NNU standard deviation terms produced very similar results, see appendix E.

6.3.3 Posterior predictive checks

Posterior predictive checks were carried out to check whether the models could have produced the observed data (methods described in section 5.2.5). Overall there was no indication from any of the posterior predictive checks of an inconsistency between predicted and replicated data.

Figure 6.1 shows the distribution of mortality rates across 1000 replicated datasets for each of the 23 neonatal networks, with data replicated according the joint posterior parameter distribution from model 1 (the basic two level model) with t-4 NNU effects. The vertical lines show the actual mortality rates observed in the data. There are no networks for which the observed data are inconsistent with the replicated data, so it is reasonable that the posterior parameter distribution could have generated the observed data. We might expect that allowance for network variation would produce replicates that were more consistent with the observed data, but this was not noticeably the case (model 4, figure 6.2. Similar plots were generated for each of the 6 models for both t-4 and Normal distributions and results were very similar across all models.

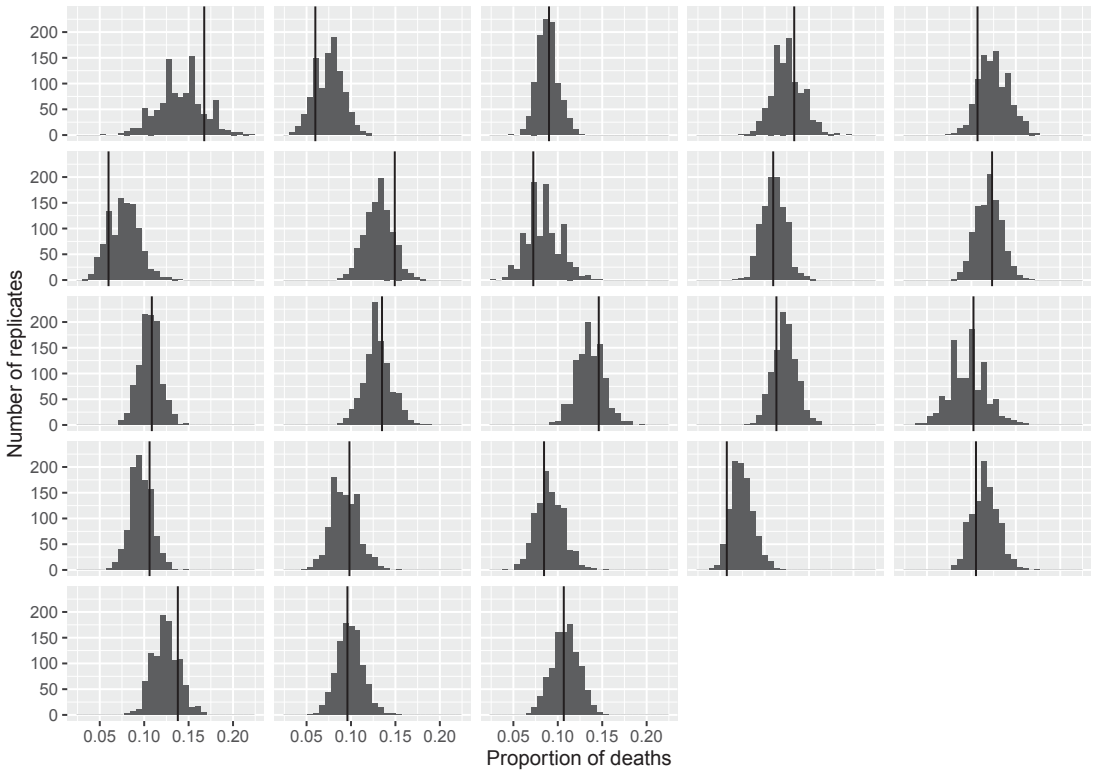


Figure 6.1: Posterior predictions from model 1 for crude mortality by neonatal network; values from data shown as vertical lines, t-4 distribution for NNU effects

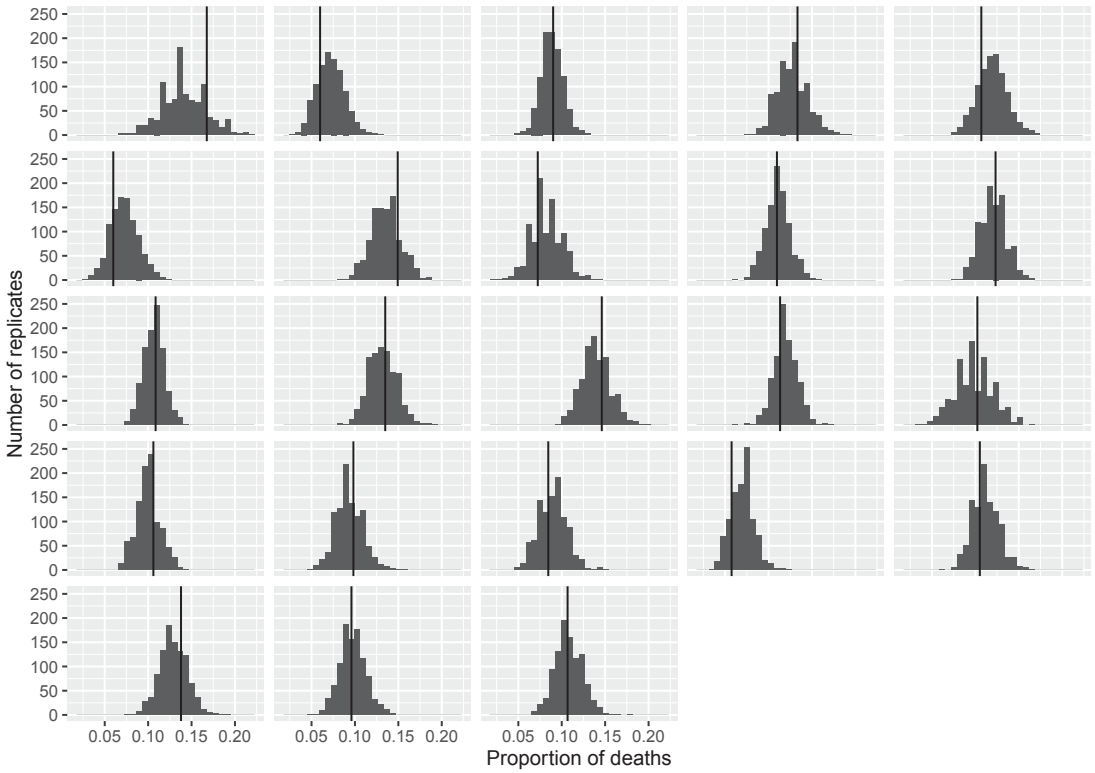


Figure 6.2: Posterior predictions from model 4 for crude mortality by neonatal network; values from data shown as vertical lines, t-4 distribution for NNU effects

As there are many NNUs, similar plots for number of deaths were generated for a random sample of 24 NNUs, and again no evidence of a disparity between the observed data and replications from the posterior distribution was found. Figure 6.3 shows this plot for model 1. There was some indication in models 1 and 4 that NNU level specific mortality was inconsistent (though not completely incompatible) between observed and replicated data; this improved after adding NNU level to the model (models 2 and 5 respectively). Figures 6.4 compares observed and replicated data across NNU levels for models 4 (without NNU level) and 5 (with NNU level).

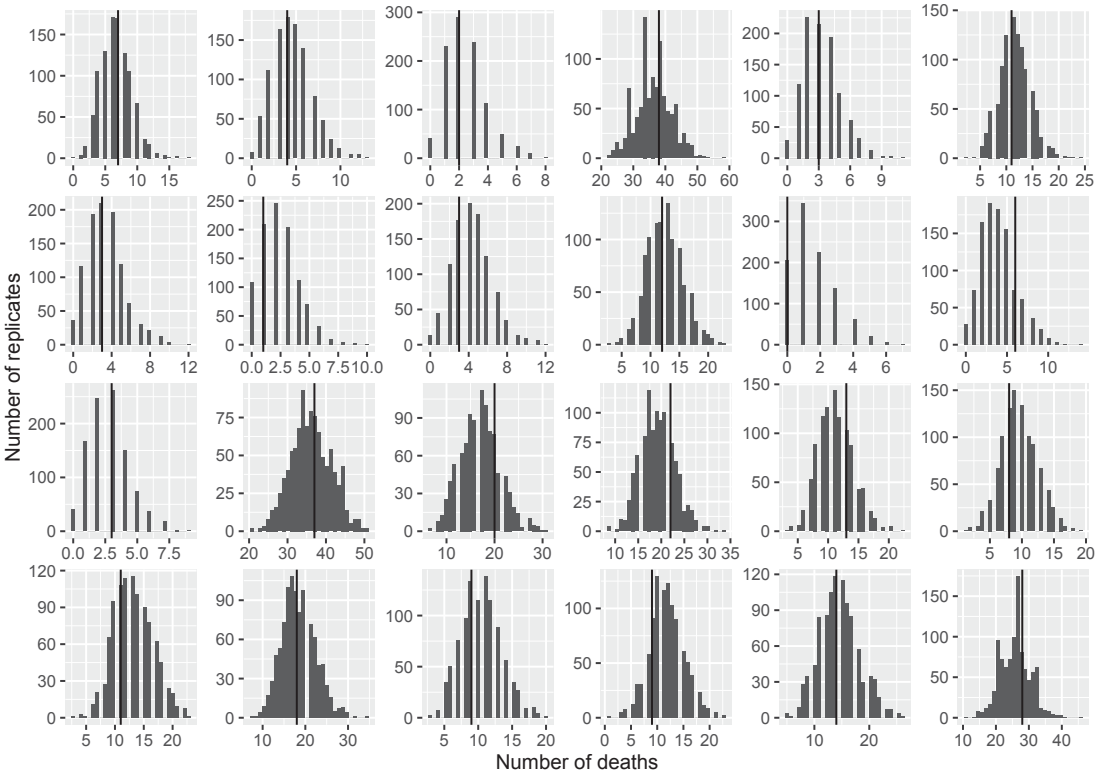


Figure 6.3: Posterior predictions from model 1 for number of deaths for 24 randomly selected NNUs with at least 50 infants; values from data shown as vertical lines, t-4 distribution for NNU effects

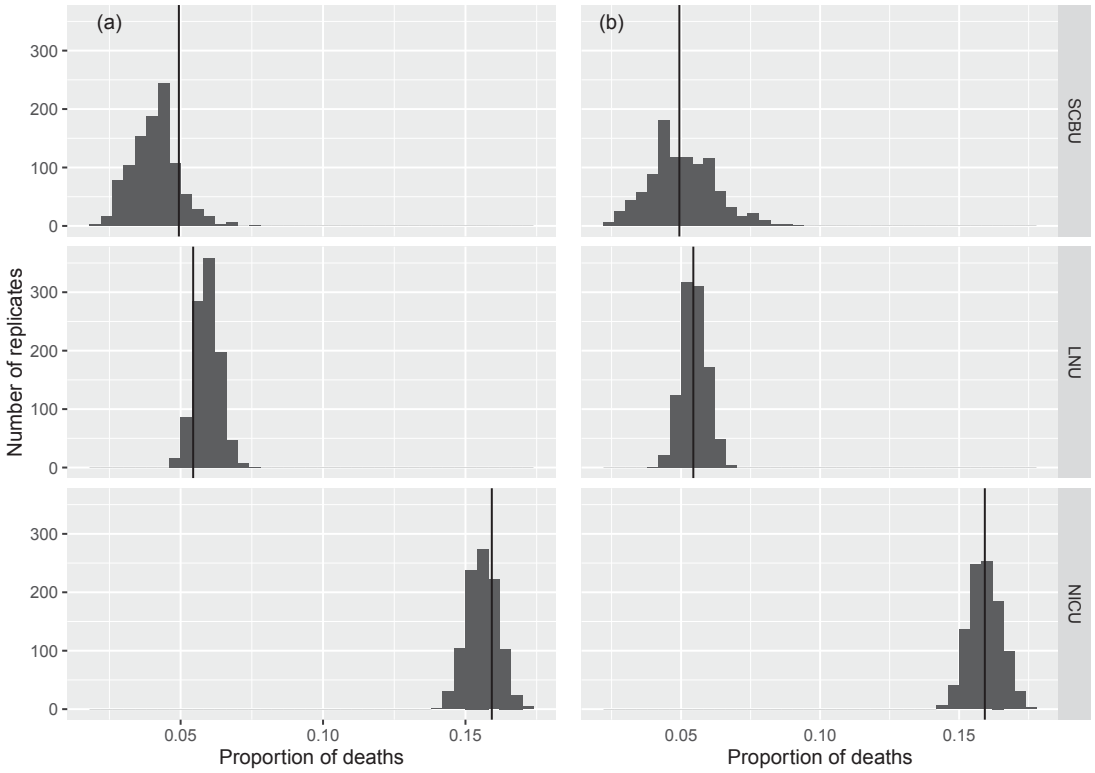


Figure 6.4: Posterior predictions from (a) model 4 and (b) model 5 for crude mortality by level; values from data shown as vertical lines, t-4 distribution for NNU effects

6.3.4 Model fit

Table 6.2 shows the Deviance Information Criterion (DIC), the deviance and the effective number of parameters for models 1 to 6 assuming a t-4 distribution for the NNU effects. There was very little difference between models; differences of less than 5 may not be indicative of model superiority (Spiegelhalter et al., 2003). Model choice will be considered in section 6.6.

	DIC	Deviance	Effective parameters
Model 1	4728	4676	52.32
Model 2	4727	4675	52.14
Model 3	4728	4676	52.00
Model 4	4728	4677	50.91
Model 5	4725	4677	48.59
Model 6	4726	4675	51.67

Table 6.2: Model performance, t-4 NNU effects U(0,5) prior

6.4 Infant and provider characteristics

Table 6.3 shows the posterior mean odds ratios (OR) and 95% credible interval (95% CrI) for the infant risk factors across all 6 models, assuming a and t distribution for the random effects respectively. These factors and their functional form were selected in chapter 3. Results were similar across all six models and between Normal at t models; results for the Normal NNU effects model are given in appendix E. Unlike the OR for covariate effects in single level models which describe the effects across the whole population, these OR have a cluster-specific interpretation; that is, they are the OR for infants in the same neonatal unit and network (see section 4.1.1). Additionally, the population is different here as only non-transferred infants are included. Nevertheless, results were fairly similar to those from the single level model developed in chapter 3, with the most notable difference being a higher OR for males (1.6 in Model 1 compared with 1.41 in table 3.5). The similarity of the results of the hierarchical and single level models is because of the low level of between-NNU and between-network variation (see section 6.5.2).

Differences in mortality across NNU level are also shown in table 6.3 for the four models which adjusted for them. Mortality is highest at SCBUs though the 95% CrI were wide and include zero (OR 95% CrI comparing with LNUs ranging from 1.44 (0.80,2.46) in model 6 to 1.53 (0.92,2.48) in model 2). Mortality was also higher for infants treated at NICUs (OR 95% CrI comparing with LNUs ranging from 1.27 (1.01,1.59) in model 3 to 1.33 (1.07,1.65) in model 5). Assuming a Normal distribution for the NNU effects

produced very similar results.

Posterior mean odds ratios and 95% credible intervals for infant and provider characteristics, t-4 NNU effects U(0,5) prior
Table 6.3:]

	Model 1		Model 2		Model 3		Model 4		Model 5		Model 6	
	OR	95% CrI	OR	95% CrI	OR	95% CrI	OR	95% CrI	OR	95% CrI	OR	95% CrI
$GA^{(1)}$	0.09	(0.05,0.15)	0.09	(0.05,0.16)	0.09	(0.05,0.16)	0.08	(0.05,0.15)	0.09	(0.05,0.15)	0.09	(0.05,0.16)
$GA^{(2)}$	3.06	(1.83,5.05)	2.93	(1.78,4.91)	2.93	(1.78,4.80)	3.14	(1.91,5.18)	3.03	(1.86,4.92)	2.97	(1.79,4.97)
BWT	0.87	(0.67,1.13)	0.88	(0.67,1.15)	0.88	(0.68,1.15)	0.86	(0.66,1.12)	0.87	(0.67,1.13)	0.88	(0.67,1.15)
BWT^2	2.03	(1.81,2.28)	2.02	(1.80,2.27)	2.02	(1.80,2.26)	2.04	(1.82,2.29)	2.02	(1.81,2.27)	2.02	(1.80,2.27)
$GA^{(1)}\times$ BWT	0.27	(0.15,0.47)	0.27	(0.16,0.48)	0.27	(0.16,0.48)	0.26	(0.15,0.46)	0.27	(0.16,0.46)	0.27	(0.15,0.48)
$GA^{(1)}\times BWT^2$	0.78	(0.68,0.89)	0.78	(0.68,0.89)	0.78	(0.68,0.89)	0.78	(0.68,0.89)	0.78	(0.68,0.89)	0.78	(0.68,0.89)
BWT $\times$ $GA^{(2)}$	1.31	(0.88,1.96)	1.29	(0.87,1.94)	1.29	(0.87,1.92)	1.33	(0.89,1.98)	1.31	(0.88,1.95)	1.30	(0.87,1.94)
Male	1.60	(1.36,1.87)	1.60	(1.36,1.87)	1.60	(1.36,1.87)	1.60	(1.37,1.88)	1.60	(1.37,1.87)	1.60	(1.36,1.87)
Steroids	0.50	(0.41,0.62)	0.50	(0.40,0.61)	0.50	(0.41,0.62)	0.50	(0.41,0.62)	0.50	(0.41,0.62)	0.50	(0.40,0.62)
SCBU*	-	-	1.53	(0.92,2.48)	1.49	(0.86,2.47)	-	-	1.47	(0.88,2.40)	1.44	(0.80,2.46)
NICU*	-	-	1.28	(1.02,1.61)	1.27	(1.01,1.59)	-	-	1.33	(1.07,1.65)	1.31	(1.05,1.63)

OR Odds ratio; CrI Credible interval
$GA^{(1)}$ and $GA^{(2)}$ denote spline terms for gestational age at birth; BWT birth weight
All continuous coefficients were standardised so OR are for a one standard deviation increase
* SCBU odds ratio comparing special care baby units to local neonatal units (LNUs); NICU odds ratio comparing neonatal intensive care units to LNUs

6.5 NNU and network effects

6.5.1 Distribution of NNU effects

The quantile-quantile plots against a Normal distribution for the NNU effects (the log odds ratios denoted by u_j) showed indication of heavy tails, even for the most complex model, model 6 (figure 6.5)(a). A slightly better fit was produced by the t distribution with 4 degrees of freedom (figure 6.5)(b), hence the t-4 models are presented as the main results.

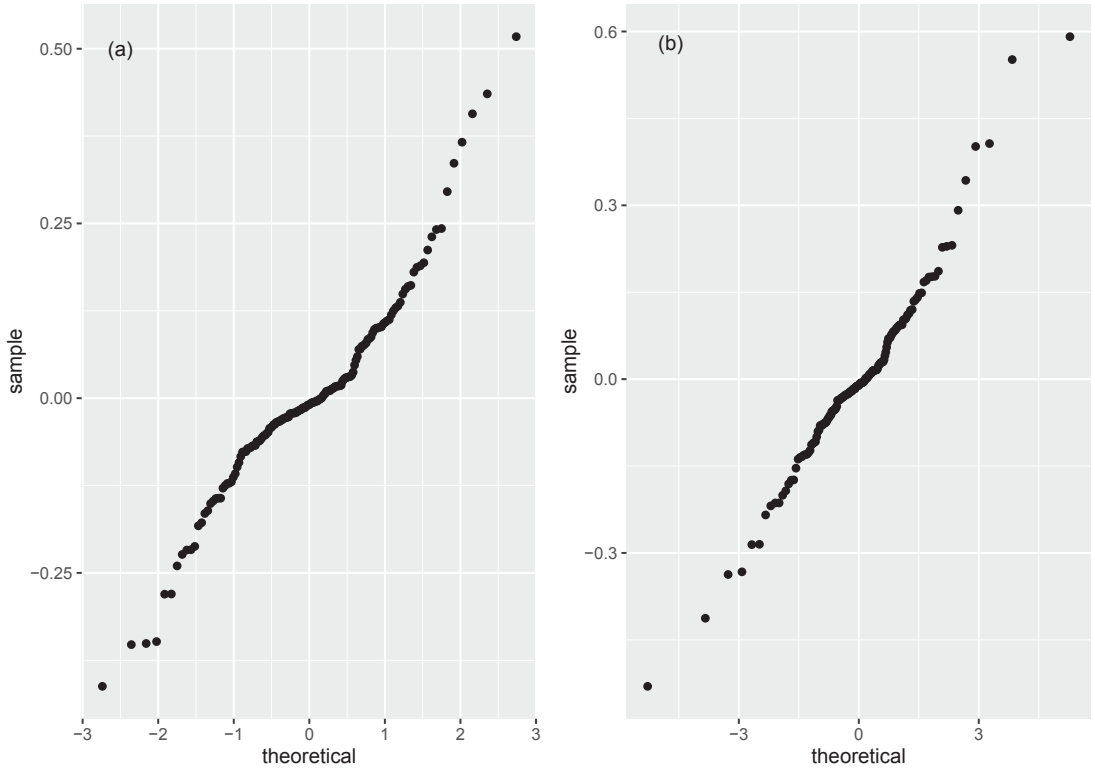


Figure 6.5: Quantile-quantile plots for posterior means from model 6, all NNUs, (a) Normal distribution and (b) t distribution with 4 degrees of freedom

6.5.2 Standard deviation of NNU and network effects

Figure 6.6 shows the posterior distribution of the between-NNU and between-network standard deviations (SD) for each of the six models, assuming the NNU effects follow a t distribution with 4 degrees of freedom. Model 1, the basic model with NNU effects but no network effects and no adjustment for NNU level , gives a between-NNU SD of 0.37 (95% CrI 0.23,0.53), corresponding to a variance partitioning coefficient (VPC) of 4% . Little of this variability was explained by the level of NNU (as σ_u for model 2 is very similar (0.36 (0.22,0.52)) but some is explained by clustering at the network level as σ_u in model 4 is 0.31 (0.13, 0.49), with a between-network SD of 0.18 (0.02, 0.,37). The VPC for NNU and network level combined in model 4 is similar at 3.8%. Once network variation is included in the model, the addition of NNU level (model 5) changes standard deviation estimates slightly with σ_u reducing to 0.27(0.07,0.46) and σ_v increasing to 0.22 (0.03,0.4); this difference is visible comparing figures 6.6 (d) and (e).

Allowing for the between-NNU SD to vary by NNU level shows a different pattern across the three levels (figure 6.6 (c) and (f)). For model 3 (the two level model without network variation) there is evidence of variation across NICUs (σ_u^{NICU} 0.41 (0.24,0.47). For SCBUs and LNUs the credible intervals are very wide, with posterior means higher for SCBUs (σ_u^{SCBU} 0.51 (0.02,1.45)) and lower for LNU (σ_u^{LNU} 0.26 (0.03,0.57). Once network variation is added, between-NICU variation is lower and more uncertain (σ_u^{NICU} 0.29 (0.07,0.48)) as some of the similarities within NICUs are explained by similarities within network. The between network variation is almost identical regardless of whether NNU variation is modelled with a single term or by NNU level.

Patterns of variation were similar for the Normal model, but the standard deviation was lower and more precise across all models, for example 0.35 (0.23,0.39) for model 1, see appendix E for corresponding plot.

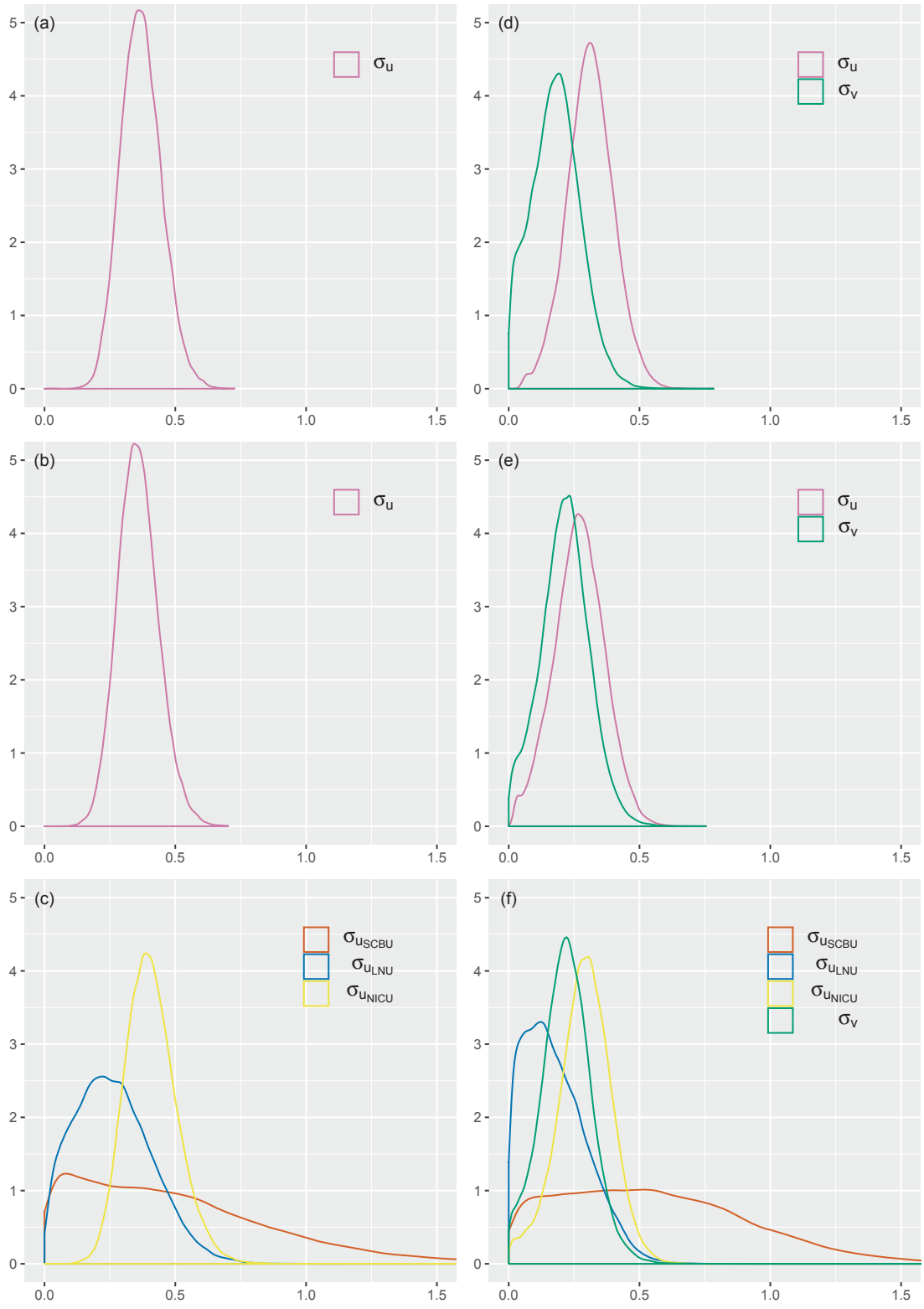


Figure 6.6: Posterior distribution provider SD terms, t-4 NNU effects. Models 1-6 shown in figures (a) to (f)

6.5.3 Comparison of NNU effects across models

Figure 6.7 shows the correlations of individual NNU effects for each pair of models, and distribution of posterior means of NNU effects by NNU level. The lowest correlation was between results for models 1 and 6 (0.78) and the highest between models 1 and 4 (0.99). Correlations within NNU level were very high (over 95% for all). The distributions of the posterior means showed markedly different patterns for all models, suggesting that whilst they may be closely correlated the values themselves might be quite different. Overall posterior means are more dispersed for models 1 to 3 which do not account for clustering at the network level, which is consistent with the higher σ_u seen in figure 6.6 (a) and (b) compared with (d) and (e). The higher the between-NNU variance, the less similar NNUs are to each other and the less we borrow information, so the shrinkage effect is smaller. This can also be seen in the scatter plots as the range is wider models 1 to 3.

As SCBUs are smaller (table 6.1), their effects are shrunk more towards zero when they are assumed to be similar to all other NNUs. When SCBUs are modelled to have their own independent variance (models 3 and 6), they are more dispersed (corresponding to the higher between-SCBU SD described earlier) and also more uncertain. Each NNU effect has an associated SD describing its uncertainty. The average SD across SCBUs was 0.35 and 0.28 in models 2 and 5, increasing to 0.55 and 0.61 in models 3 and 6 (table 6.4). In contrast, for LNUs the effects are shrunk more when LNUs have an independent variance (models 3 and 6), so they are more similar to each other than NNUs are to each other in general, under the common variance model. The shrinkage effect for LNUs is larger when network variation is also included in the model. Including network effects also results in more precise estimation of effects (table 6.4). NICUs, being the type of NNU with the highest volumes, experience the least shrinkage, particularly in models 1-3 where variance at the network level is not modelled.

Table 6.4: Posterior mean standard deviation for NNU effects, averaged over NNU, t-4 distribution for NNU effects, U(0,5) prior

	All	SCBU	LNU	NICU
Model 1	0.31	0.36	0.32	0.24
Model 2	0.30	0.35	0.31	0.24
Model 3	0.33	0.55	0.26	0.25
Model 4	0.28	0.31	0.28	0.23
Model 5	0.25	0.28	0.25	0.22
Model 6	0.31	0.61	0.20	0.24

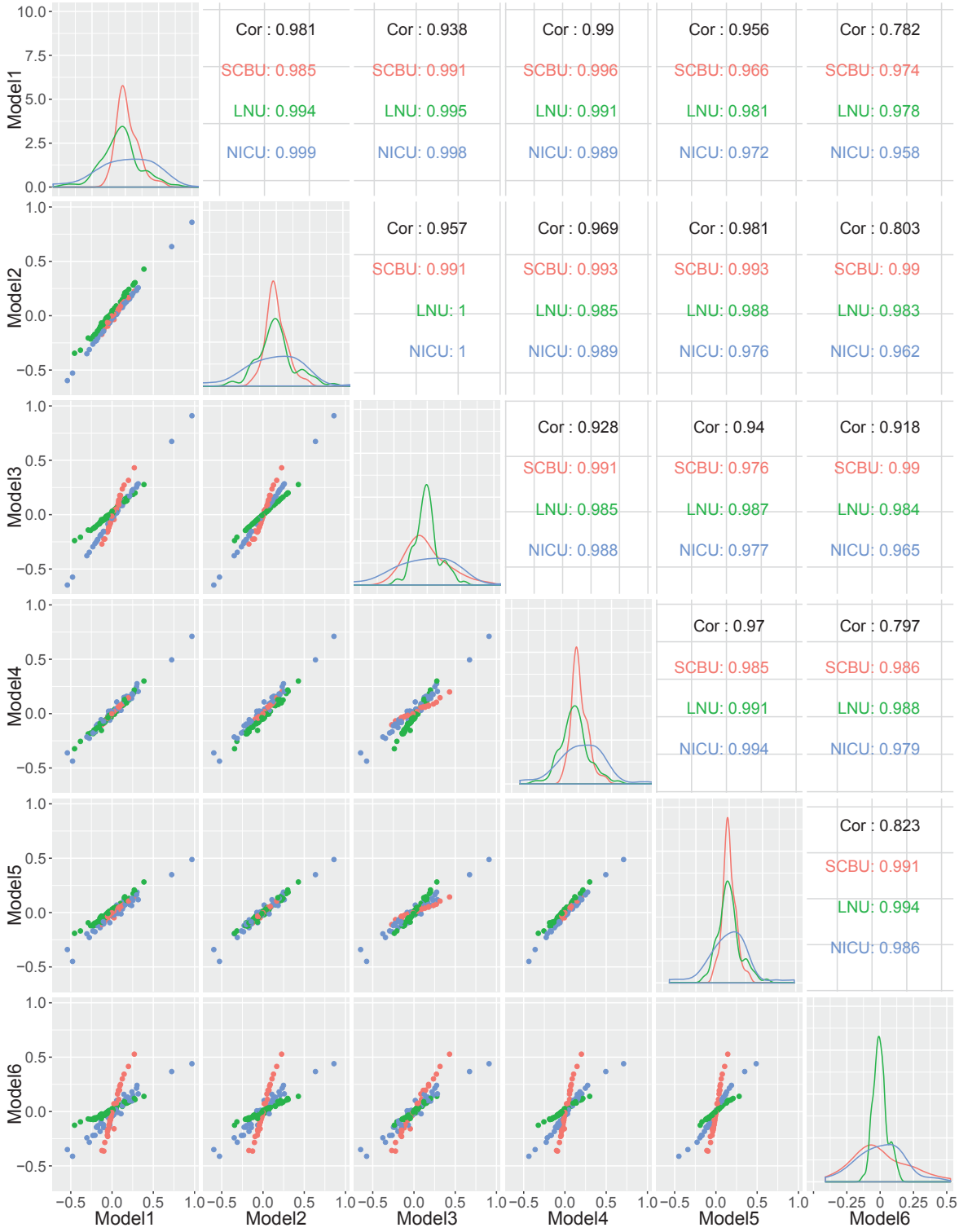


Figure 6.7: Correlation of NNU effects across models, t-4 distribution for NNU effects

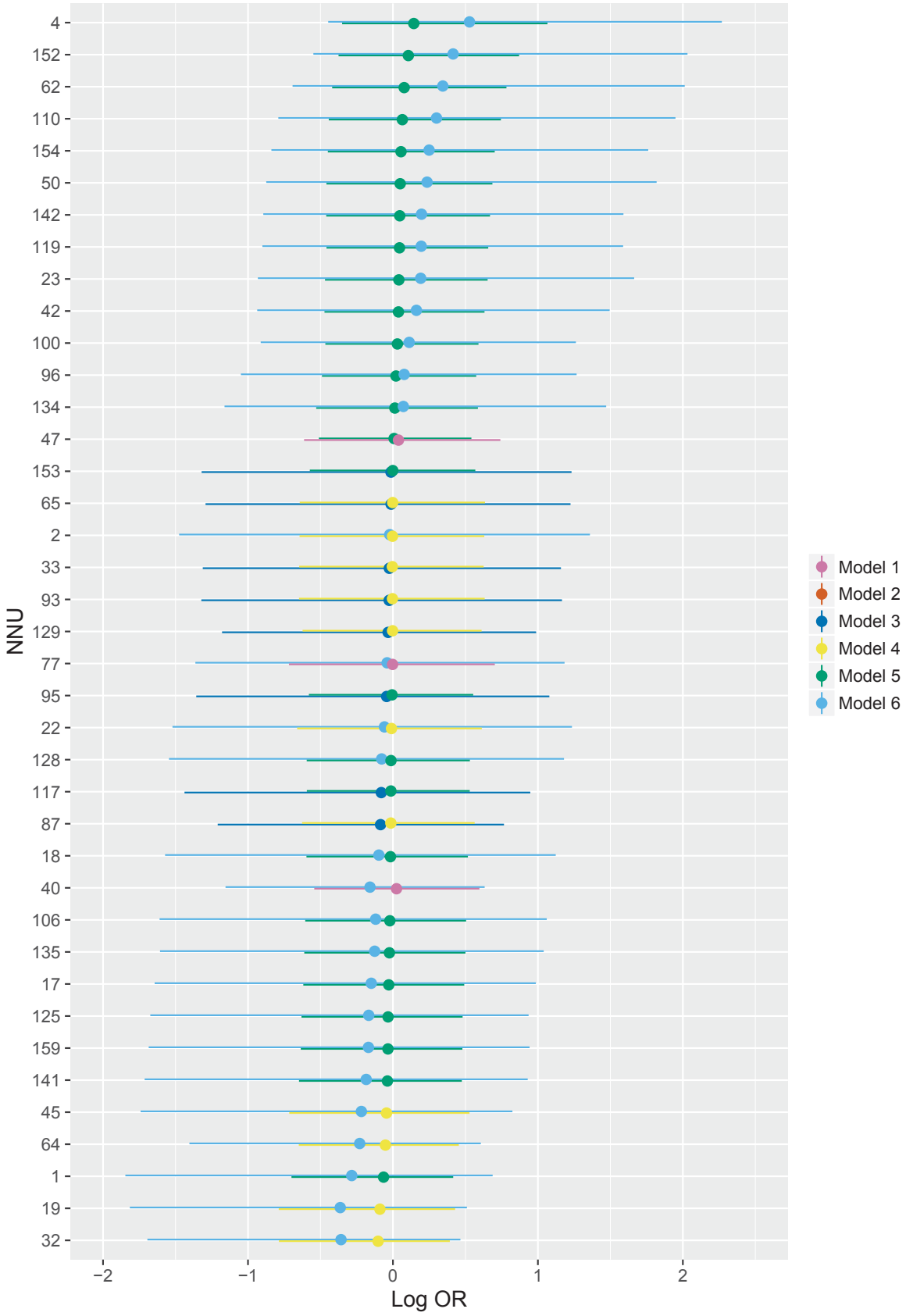


Figure 6.8: Caterpillar plot for SCBUs, showing smallest and largest effects across models, t-4 NNU effects

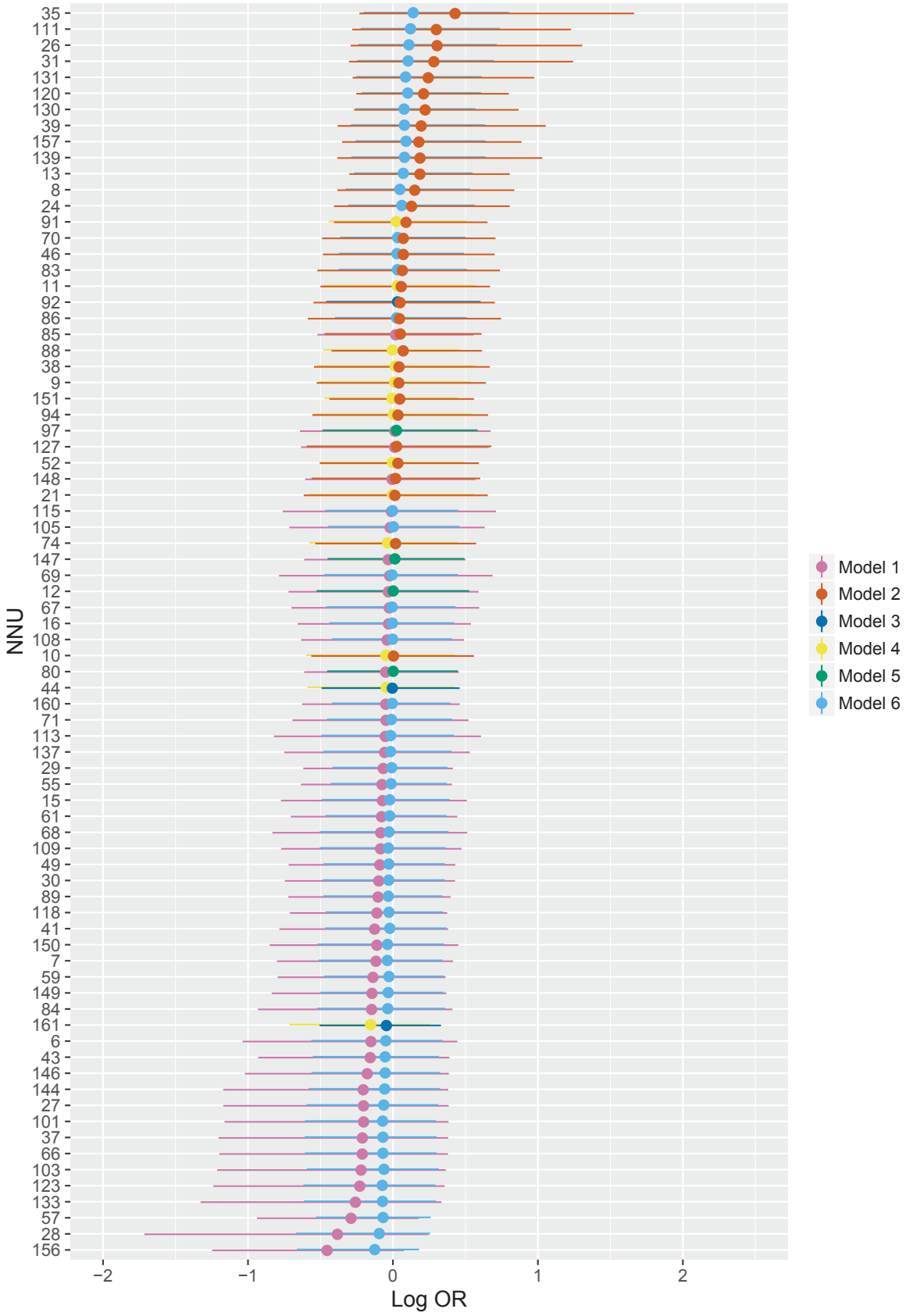


Figure 6.9: Caterpillar plot for LNUs, showing smallest and largest effects across models, t-4 NNU effects

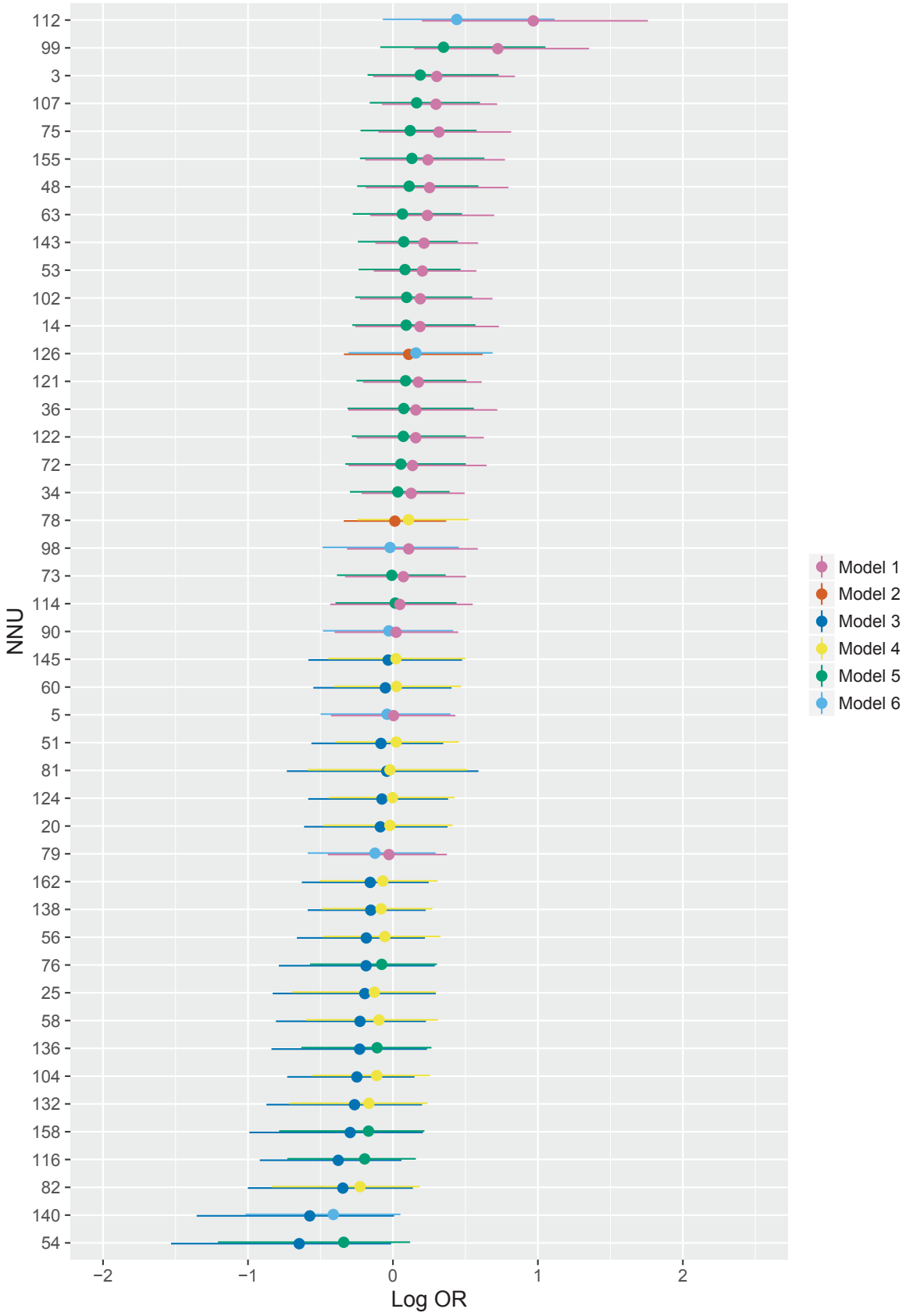


Figure 6.10: Caterpillar plot for NICUs, showing smallest and largest effects across models, t-4 NNU effects

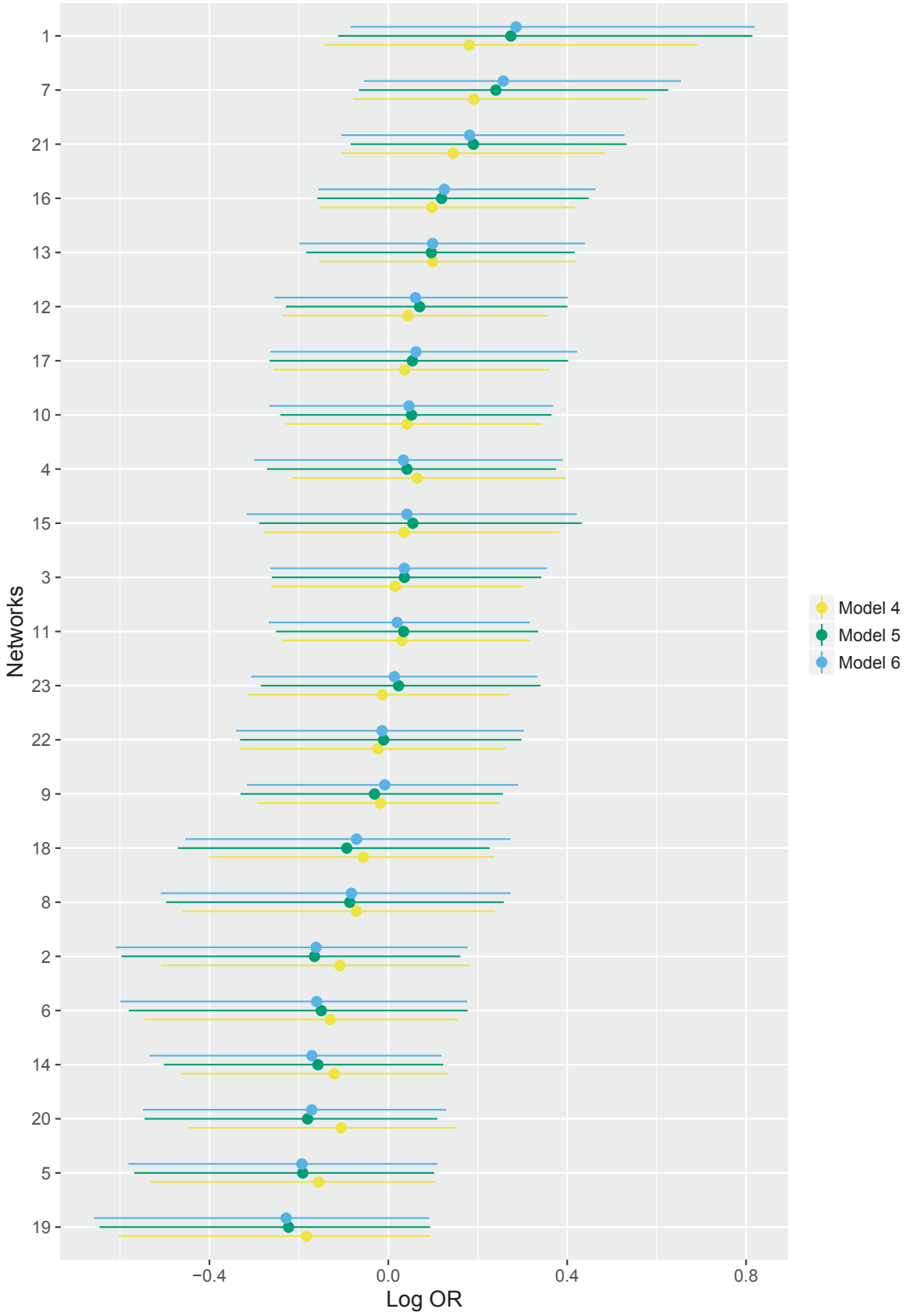


Figure 6.11: Caterpillar plot for networks, showing smallest and largest effects across models, t-4 NNU effects

Figures 6.8 to 6.11 show plots comparing SCBUs, LNUs, NICUs and neonatal networks respectively. They show the smallest and largest posterior mean log OR produced by the models with 95% CrI. The NNU effects are ranked to better illustrate patterns, though as discussed in chapter 4 no inferences can be drawn regarding whether one NNU is better or worse than another based on the rank order. These plots are sometimes called *caterpillar plots*. The reduced shrinkage of SCBU effects produced by model 6 is evident in figure 6.8 as model 6 produces the highest estimates for SCBUs with high log OR, and lowest estimates for SCBUs with the lowest log OR, thus the overall distribution of posterior means is more dispersed. Model 6 also produces much wider credible intervals for SCBU effects which contain the credible interval of the least extreme result. This increased uncertainty is in line with the higher average SD seen in table 6.4. The smallest absolute effects (where they are noticeably different from the largest absolute effects) were from models 4 and 5, which from figure 6.7 produced the least dispersed posterior means.

For the majority of LNUs there was little difference in the smallest and largest effects in the context of the uncertainty of the estimates (figure 6.9). Where there were differences, the highest log ORs were generally produced by model 2 and the lowest by model 6, which is consistent with the patterns seen in figure 6.7. Despite the uncertainty, differences may be meaningful when we consider the differences in terms of the probability of exceeding a certain threshold. For example for NNU 73, model 2 estimates the posterior probability that the OR exceeds 1.2 (equating to a log OR of 0.18) as 65%, while model 6 gives an estimate of 31%.

Posterior means of NICU effects are not shrunk as much as smaller NNUs , so in figure 6.10 there are fewer NNU effects close to the zero line. Whereas for SCBUs and LNUs the smallest and largest effects was negligible for NNUs in the middle (those with effects near zero) this was not the case for NICUs. For example, the probability of the OR for NNU 73 exceeding 1.2 was 28% in model 1 but 12% in model 5. Models 4 and 5 produced estimates closest to zero. Model 1 tended to produce the highest estimates and model 3 the lowest. Model 1 is not adjusted for NNU level, therefore the increased mortality seen in NICUs overall as shown by the ORs in 6.4 is included in the individual NNU effects. Credible intervals for NICU effects are much narrower than for SCBU and LNU effects, consistent with the lower average SD seen in table 6.4.

Figure 6.11 shows that there is little variation in network effects across models given the overall uncertainty. Estimates from model 4 (which does not include NNU level) produced NNU effects with slightly smaller magnitude (consistent with the lower estimate of between network variation seen in figure 6.6 (d)), but differences were small.

Corresponding plots for Normal NNU effects showed a similar pattern and are pro-

vided in appendix E.

6.6 Discussion

6.6.1 Summary

In this chapter I presented results from the analysis of non-transferred infants, applying the methods described in section 5.2. There were differences in mortality between NNU levels beyond that attributable to the included risk factors, with both SCBUs and NICUs having higher mortality rates than LNU. Infants treated at the same NNU and the same network had similar outcomes, though this accounted for only 4% of the overall variance as estimated by the VPC. It is possible that variation is higher for SCBUs and NICUs compared with LNU, though the evidence for this was weak as standard deviation estimates were very uncertain. Estimates of individual NNU effects were also subject to a lot of uncertainty, but differences in results across models were meaningful for some NNUs. Assuming a t distribution with 4 degrees of freedom appeared to fit the data better than the standard approach of Normal random effects. There was no indication of a discrepancy between the model and the data for any of the models based on posterior predictive checks and there was very little difference in DIC. However there were some differences regarding variation across NNU, both with overall measures and for individual NNU results that may affect model choice, which I shall discuss further in this section.

6.6.2 Model checking

I used a variety of convergence checks, which was important as I found that the initial assessment from trace plots and the Brooks Gelman Rubin statistic often indicated convergence when the other tests did not. There was poor mixing for some parameters despite taking steps to prevent this by centring and standardising covariates, and specifying the model using hierarchical centring parametrisation (see section 5.1). I used the default samplers chosen by OpenBUGS (different samplers are chosen automatically for different parameters) so it is possible that mixing could be improved and adequate effective sample size achieved more quickly by altering the sampler specification. For consistency with this chapter I continue to use the default samplers in chapter 7 but would consider alternatives in any future analysis.

Mixing was particularly poor for the intercept term where multiple variance parameters were being estimated in models 3 to 6. Figure 6.6 shows that with the exception

of models 1 and 2, some of the posterior mass for the standard deviation terms is very close to zero. Were the random effects simply nuisance parameters, as is often the case in hierarchical models, this might support their omission. Similarly there was little evidence of model superiority from the DIC, supporting selection of the simplest model if the random effects were nuisance parameters. Here it is the random effects themselves that are of interest so model choice should be driven by the inferences we can draw from these parameters. However the lack of difference in DIC does tell us that the differences between networks and NNU levels are not substantial enough to improve the fit of the model. This could be a useful when considering the potential impact of clustering in other studies of preterm mortality.

I found that a t-4 distribution for the NNU effects gave a better fit than the Normal distribution. The latter is often used as the default specification for random effects. There was no further indication of poor fit or asymmetry so I did not consider more flexible alternatives or skewed distributions mentioned in section 4.1. However, sensitivity to the degrees of freedom of the t distribution was not investigated. The t distribution provides a better fit as it accommodates NNUs with extreme results as observed in figure 6.5. While the NNUs which appear outliers may indeed be so, and this warrants further investigation, this is not an aim of this thesis. As described in section 4.4, for many NNUs there is limited power to confidently determine outlier status, thus a fair assessment cannot be made.

Results were robust to whether the priors for the NNU and network standard deviations were set as U(0,5) or U(0,100). Both are vague priors; for further discussion on the choice of prior see section 5.2. A limitation of the analysis is that I did not consider any informative priors. There was a lot of uncertainty around the NNU effects and it is possible that this could be improved by using a more informative prior for the between-NNU standard deviation. If NNU effects were to be regularly assessed as part of routine monitoring, understanding of variations in mortality across NNUs would develop. This could enable a selection of informative priors to be specified based on previous patterns.

6.6.3 Infant and provider characteristics

There was some indication that mortality was highest in SCBUs. The difference had a wide credible interval but the direction of the association is consistent with several studies advocating centralisation of neonatal care due to poorer outcomes at less experienced NNUs (Binder et al., 2011; Neto, 2006; Phibbs et al., 2007). However mortality was also higher at NICUs than at LNUs in this study. As described in chapter 1 one of the aims of neonatal networks is that where possible infants requiring the highest levels of care should be treated at NICUs. Therefore in general we would expect infants

treated at NICUs to be sicker than those treated at LNUs, resulting in increased ICU mortality. Indeed, Watson et al. (2014) found the benefits of tertiary care on mortality were underestimated in a standard regression analysis compared with an instrumental variable analysis which attempts to estimate causal effects as if infants were randomised to NNU level. It should be noted that while the direction of effect in the standard regression analysis still showed lower mortality in NICUs in contrast to the findings here, the population was slightly different as transferred infants were included and assigned to the NNU of birth. This resulted in a very different patient profile, with 6% mortality at NICUs compared with the 16% seen here. Based on the findings of Watson et al. (2014) which used the same data source, it seems unlikely that the difference in mortality between LNUs and NICUs represents a difference in quality of care.

There are two factors contributing to this difference. Firstly, infants at NICUs are more sick to begin with. This is plausible as, unlike LNUs, the NICU population includes sicker infants who had a planned delivery or in-utero transfer there. Secondly, as care progresses infants who do not require transfer are likely to be sicker ones at NICUs and less sick ones at LNUs. For LNUs non-transferred infants are those that did not require NICU care (though they also include those who died so soon after birth that they could not be transferred). While SBCU do not in general provide intensive care (which many infants born at less than 32 weeks will need), LNU often do, with considerable variation in the volume provided across LNU as described in chapter 2. For NICUs, non-transferred infants include the local population for whom the NICU is the nearest NNU, but also those that did not recover to have a subsequent step-down transfer to their nearest NNU.

Some differences in population across NNU levels would be captured by the risk adjustment model as the distribution of risk factors differed across levels of NNU (table 6.1). There may be other infant characteristics which determine the place of delivery (and are therefore known at birth) which would improve the risk adjustment model but these were not revealed by the risk factor selection process described in chapter 3. While such factors may not be such strong predictors of mortality, they could be strong confounders if they dictate where an infant is cared for. Further work is required to examine the reasons determining NICU delivery and establish whether such information can be included in the risk adjustment model.

In this thesis the primary purpose of including NNU level is not to estimate a causal effect of NNU level. The purpose of including NNU level in the model is to indicate whether mortality at a particular NNU is higher or lower than would be expected for infants of the same clinical characteristics treated at NNUs of the same level. Given the differences in the care provided at each NNU level described in chapters 1 and 2, and the differences in mortality described here, this seems a sensible approach.

6.6.4 NNU and network effects

Variation in mortality across NNUs accounted for 4% of the total variation as estimated by the VPC for model 1. Using this method of calculation of the VPC we consider the binary variable as a manifestation of an underlying continuous variable (see section 5.2.6), which gives the individual-level variance on the continuous scale. Other methods exist which produce VPC estimates based on the probability of the response, and are dependent on the individual-level covariates (see Goldstein et al. (2002) for further details). It is reasonable to infer that differences between NNUs are not responsible for large variations in mortality, though our inferences are on the continuous scale of the latent variable (the propensity to die) rather than on the probability scale. Some of this variation was attributable to neonatal networks; this was more apparent after adjusting for NNU level. Networks consist of a combination of levels of NNU, therefore σ_v which measures the extent to which NNUs in the same network are similar, will be dampened by the the differences between NNU levels until this is adjusted for. Network variation is of interest in itself as many commissioning, operational and policy decisions are made at the network level. Although NNUs and networks combined did not account for any more variation than NNUs alone, it is important to adjust for network variation when estimating NNU effects to avoid mistakenly attributing variation at the wrong level. In addition, accounting for clustering by network gives slightly higher precision in the NNU effects (table 6.4).

There was little evidence that between-NNU variation differed by type of NNU: while posterior means differed, figure 6.6 shows that effects are very uncertain and did not improve the fit of the model.An argument for using hierarchical models in provider profiling is to enable the borrowing of information for smaller providers. We cannot do this if we allow the between-NNU variance to differ across NNU levels because all SCBUs have few admissions in this gestational age group. If SCBUs can only borrow information from each other, the resulting estimates are extremely uncertain, as shown in figure 6.8 and table 6.4. Therefore I will not allow for level-specific variance in chapter 7. However it is interesting to note the possibility of higher variation between SCBUs. This could be an artefact of small sample sizes, but could also represent the varying success with which infants are stabilised and transferred out for further care at lower level NNUs (Marlow et al., 2014). A larger sample size would be required to investigate this further.

Provider profiling analyses comparing different methods (for example those described in section 4.2) often focus on changes in rank, particularly for the providers at the top and bottom. As discussed in section 4.4.2, the number of infants at each NNU is variable and overall fairly small. This results in variability in shrinkage with smaller NNUs being shrunk closer to the mean, and low power to detect outliers in general,

so in this thesis I have focussed on estimation of individual NNU effects rather than comparison among them or assessment of outliers. NNU effects from all models were closely correlated within NNU level but not overall, suggesting that ranking results (though not a focus of this thesis) would give similar results regardless of the model used provided this was done separately for NNU levels. However the distributions of posterior means were quite different across models, so there are differences in the actual estimates, which are of interest here. The caterpillar plots showed that the models producing the highest and lowest effects differ by NNU level, therefore the choice of model will impact some NNUs differently from others. For many NNUs, particularly the smaller ones, the differences between models are not notable in the context of uncertainty of the estimates. However this is not always the case, particularly when we consider NNU effects in terms of probability of exceeding a threshold, a measure which is often considered more interpretable than a confidence interval or credible interval (Normand et al., 1997). There were different amounts of shrinkage for different NNU levels even when the same variance was assumed across levels. This is to be expected as patient volume differs by NNU level, re-iterating the difficulty of directly comparing results for NNUs at different levels. It also means that we cannot draw the same conclusions about a SCBU with an average result as we can for a NICU: for example from model 6, SCBU 65 and NICU 145 both have OR very close to 1 (0.99 and 1.01 respectively) but their credible intervals support very different ranges (SCBU 65 (0.23,4.08) and NICU 145 (0.60,1.68)).

The approach taken here is based on existing Bayesian hierarchical methods for provider profiling. As described in chapter 4, these models have been applied in neonatal care but I am not aware of any other studies specifically examining NNU and network effects in England or another setting with a similar structure. Therefore I am unable to draw comparisons with other studies regarding these effects, particularly with regard to how they are affected by modelling strategies as was the main focus of the investigation here. MacNab et al. (2005) investigated variation in NICU mortality in Canada; while this was not in a networked setting they did assess both Normal and t-distributions (with 2 to 5 degrees of freedom) for the random effects. Near identical parameter estimates were obtained with no indications of poor fit though there were only 17 NNUs, so departures from distributional assumptions may not have been clear. They found conclusions to be robust to a variety of hyperpriors for the standard deviations, applying a wider selection than the Uniform priors applied here. Further sensitivity analysis would help make the conclusions drawn in this thesis more robust.

6.6.5 Clinical summary

From a clinical perspective, there does appear to be variation in mortality between NNUs and networks but it is low, and accounts for a small proportion of the overall variation. For each NNU we obtain an ORs comparing the mortality at that NNU to that for similar infants (based on gestational age, birth weight, sex and antenatal steroid administration) in similar NNUs (based on neonatal network and NNU level, depending on the model). However, these are subject to high levels of uncertainty so it is difficult to say conclusively whether the outcomes are better or worse than expected. While the lack of variation is reassuring for those managing services, is also means that the results do not provide strong evidence of which NNUs may require additional attention or resources. Tail probabilities for the OR (as given here as examples) may be a useful measure in practice but suitable thresholds for the OR must be chosen.

There are differences between NNU levels in the non-transferred populations that go beyond that described by the main clinical risk factors. Some of the difference may be due to factors known at birth determining place of delivery, while some is due to selection bias by examining only non-transferred infants. While bias from selecting only non-transferred infants for a research cohort is well-understood, this indicates that the selection effects differ across NNU levels.

6.6.6 Implications for the thesis

One of the aims of this chapter was to develop a suitable model on which to base the multiple membership analysis which follows in chapter 7. From the findings of this chapter I will include networks as a third level of variation in the model and allow the mean of NNU effects to be specific to NNU level, but with a common variance (model 5). While networks do not account for any more variation than NNU alone, the network effects are themselves of interest. Furthermore, it means the variance will be correctly attributed between NNUs and networks, and gives a slight increase in precision of NNU effects. Adjusting for NNU level ensures that NNUs are compared with the relevant average. This is particularly important given the substantial differences in infants treated at different NNU levels which goes beyond the infant characteristics controlled for. In addition the mixture of NNU levels within each network appears to dampen the between-network variation, though the effect on individual network results was negligible. There was little evidence for allowing different variance between NNU levels. Doing so limits the extent to which we can borrow strength to estimate effects at smaller providers and may result in poor estimates if the variances are unstable. Model 5 also produces the least extreme results for many SCBUs and NICUs (figures 6.8 to 6.10), lowest uncertainty in NNU effects and the least variation in uncertainty across

NNU levels (6.4).

In chapter 7 I build on model 5 to include transferred infants by applying multiple membership models as described in section 5.4. Results from model 5 in this chapter are presented for a single NNU as part of a case study in chapter 8.

Chapter 7

Results II: Including transferred infants

In this chapter I present the results of applying the methods described in sections 5.3 and 5.4.

Multiple membership (MM) models have been proposed as way to handle postnatal transfers in neonatal provider profiling analyses. An infant's care is not attributed to a single neonatal unit (NNU). Instead it is shared between all NNUs treating the infant using weights. In section 7.2 I present a simulation study to investigate the effect of different MM weightings on the between-provider standard deviation and individual provider effects under different data scenarios. Section 7.4 builds on the results of chapter 6, including transferred infants in the analysis by weighting the NNU effects using multiple membership models. Infant and provider characteristics (section 7.4.4), and provider effects (section 7.4.5) are compared across six different multiple membership weighting schemes. In section 7.5 I summarise the findings from the chapter and consider the practicalities of applying these methods both within neonatal care and more widely.

7.1 Methods recap for simulation study

This section is a brief summary of the methods underlying the results for this section; full details were given in chapter 5.3. The aim of the simulation study was to investigate the differences in variance parameters and provider effects under a variety of data scenarios and analysis models for multiple membership, with a view to informing the analysis of the neonatal data where the parameters generating the data are unknown. The simulation

was done prior to the analysis presented in 6 and so does not reflect some of the insights from that chapter, such as modelling the provider effects with a t-distribution rather than Normal, and some of the parameters do not equate with what was observed in the data.

A two-level structure was considered (individuals within providers). Data were simulated for 15,000 individuals across 100 providers, with 25% of individuals transferred once. Provider-specific log odds of mortality were drawn from a Normal distribution with a mean of -2.2 and standard deviation 0.4, corresponding to 95% of provider mortality rates lying between 5% to 20%. No infant or provider characteristics were included. Data were generated using three scenarios with different MM weightings: proportional to length of stay (denoted *LOS*), moderate emphasis on the early part of stay (*Beta2*), and a strong emphasis on the early part of stay (*Beta3*). For each of the three scenarios I simulated 200 datasets (600 datasets in total).These weightings, along with allocation to the initial provider only (denoted *First*) were used to analyse the data, resulting in 12 sets of results in total. For reference, weightings are summarised in table 7.1, which also describes two weights only used in the analysis of neonatal data in section 7.4. Bayesian Markov Chain Monte Carlo (MCMC) methods were used to estimate the models. Average posterior means, empirical standard errors (SE), percentage bias, mean square error (MSE) and coverage for the standard deviation and intercept were calculated and the average bias, MSE and coverage for the provider effects (see section 5.3.5 for definitions of these quantities). The Deviance Information Criterion (DIC) was calculated and used to compare models.

Table 7.1: Weightings for multiple membership models for simulated and neonatal data

Label	Description
LOS	The outcome is attributed in proportion to the length of stay
Beta2	The outcome is attributed using the Beta(1,2) weight
Beta3	The outcome is attributed using the Beta(1,3) weight
First	The outcome is attributed to the first provider only
WLOS (neonatal data only)	The outcome is attributed according to length of stay weighted by the type of care received, using recommended nurse staffing ratios
Equal (neonatal data only)	The outcome is attributed equally to all NNUs caring for the infant, regardless of length of stay or intensity of care

7.2 Results from simulation study

Results from analysing simulated data are shown in table 7.2. The mean intercept (the average log odds of death) was correctly estimated as -2.2 for all scenarios and all analysis models, with low bias (<1% for all cases).

Between provider variation The posterior mean between-provider standard deviation (σ_u) was underestimated using the *First* analysis weighting, with the largest downward bias of 36% for the *LOS* scenario, reducing to 15% for the *Beta3* scenario. The *First* analysis yielded poor coverage for the σ_u (as to be expected when bias is high), with the 95% credible interval (95% CrI) including the true value in only 3% of iterations for *LOS* scenario increasing to 73% of iterations for the *Beta3* scenario. The average σ_u was similar for models which do not ignore the MM structure, with estimates ranging from 0.36 to 0.42, compared with the true value of 0.4. The average σ_u was always lower when the analysis weighting was more "uneven" than the scenario weighting but differences between MM models were small. A more even analysis weighting than the scenario weighting did not consistently result in a higher σ_u. Using the same weighting to generate and analyse the data did not consistently minimise the bias or MSE for σ_u.

Provider effects There was a large downward bias in the estimation of the posterior mean provider effects (around 50% for matching scenario and analysis weightings) due to shrinking of the estimates towards the mean of the random effects distribution. In all scenarios, matching scenario and analysis weightings gave the smallest average bias and MSE for the provider effects, but again the differences between MM models were very small. The *First* analysis gave the most biased but most precise estimates (as measured by the average posterior SD of provider effects), resulting in poorer coverage than the MM models.

DIC Matching scenario and analysis weightings produced the lowest average Deviance Information Criterion (DIC). However the proportion of iterations where the difference in DIC between the scenario and analysis weightings was at least five was variable (79% for *LOS*, 19% for *Beta2* and 40% for *Beta3*), with the variation in percentages reflecting the similarity of alternative models.

			Analysis weighting			
Scenario weighting	Parameter		*LOS*	*Beta2*	*Beta3*	*First*
	σ_u	Mean	0.415	0.389	0.364	0.257
		SE	0.038	0.037	0.037	0.040
		% Bias	3.9	-2.8	-9.1	-35.8
		MSE	0.002	0.002	0.003	0.022
		% Coverage	99.0	96.5	88.5	3.0
LOS	α_0	Mean	-2.216	-2.213	-2.209	-2.193
		SE	0.029	0.029	0.029	0.029
		% Bias	0.7	0.6	0.4	-0.3
		MSE	0.001	0.001	0.001	0.001
		% Coverage	100.0	100.0	100.0	98.5
	u_j	Ave % Bias	-42.8	-46.3	-51.5	-72.1
		Ave MSE	0.060	0.063	0.066	0.094
		Ave Coverage	95.5	93.8	92.2	76.9
		Ave SD	0.255	0.243	0.233	0.187
	DIC	Mean	9828	9838	9849	9897
		% correctly identified	78.5	–	–	–
	σ_u	Mean	0.414	0.414	0.402	0.314
		SE	0.035	0.033	0.032	0.032
		% Bias	3.4	3.5	0.4	-21.6
		MSE	0.001	0.001	0.001	0.008
		% Coverage	99.0	99.5	99.0	42.0
Beta2	α_0	Mean	-2.212	-2.215	-2.215	-2.201
		SE	0.028	0.028	0.028	0.027
		% Bias	0.6	0.7	0.7	0.1
		MSE	0.001	0.001	0.001	0.001
		% Coverage	100.0	100.0	100.0	100.0
	u_j	Ave % Bias	-58.2	-55.2	-56.0	-65.2
		Ave MSE	0.059	0.057	0.057	0.073
		Ave Coverage	95.6	95.7	95.1	86.9
		Ave SD	0.254	0.249	0.243	0.207
	DIC	Mean	9851	9841	9843	9886
		% correctly identified	–	18.5	–	–
	σ_u	Mean	0.407	0.420	0.414	0.338
		SE	0.036	0.034	0.034	0.032
		% Bias	1.7	5.1	3.5	-15.4
		MSE	0.001	0.002	0.001	0.005
		% Coverage	99.0	99.0	99.0	72.5
Beta3	α_0	Mean	-2.205	-2.212	-2.213	-2.202
		SE	0.026	0.026	0.026	0.026
		% Bias	0.2	0.5	0.6	0.1
		MSE	0.001	0.001	0.001	0.001
		% Coverage	100.0	100.0	100.0	100.0
	u_j	Ave % Bias	-55.2	-48.7	-48.7	-61.4
		Ave MSE	0.061	0.057	0.056	0.066
		Ave Coverage	95.5	96.0	95.6	89.4
		Ave SD	0.252	0.251	0.246	0.214
	DIC	Mean	9888	9867	9864	9899
		% correctly identified	–	–	40.0	–

Table 7.2: Results from simulation study comparing four analysis models of three different multiple membership scenarios

In summary the simulation study which showed that ignoring multiple membership structure and assigning individuals to a single provider resulted in underestimating the provider variance and poorer coverage of the variance and provider effects. Biases were larger when the weights generating the data were more evenly distributed than those used to analyse the data. In comparison, biases from assuming an incorrect MM weight were minimal. There was some indication of variance underestimation when a more uneven attribution was assumed but not the converse. Estimates of provider effects themselves were less strongly affected than the provider variance, with any biases overshadowed by shrinkage effects. While the average DIC was lowest for correct analysis models, the DIC had limited utility for model selection. These results will be discussed in section 7.5.

7.3 Methods recap for analysis of neonatal data

This section is a brief summary of the methods underlying the results for this section; full details were given in chapter 5.4. Following from the results of chapter 6, three-level Bayesian hierarchical logistic regression models (infants within NNUs within networks) were used to analyse the data, estimated using MCMC. The response variable was death before discharge from neonatal care. The following infant risk factors were adjusted for: gestational age (as a restricted cubic spline with 3 knots), birth weight (as linear and quadratic), all interactions between gestation and birth weight involving linear terms, sex and whether the mother received antenatal steroids. The level of NNU was included as a provider-level characteristic. The population was all singleton infants born below 32 weeks gestation between 2011 and 2013 and admitted to an NNU contributing data to the National Neonatal Research Database (NNRD) including those transferred during the course of neonatal care. A t distribution with four degrees of freedom was assumed for the NNU effects, with a common variance across NNU levels. This corresponds to model 5 in section 6.1.

Transferred infants were included by applying MM models. The weightings applied were the four described for the simulation study (*LOS*, *Beta2*, *Beta3* and *First*) together with weighting the length of stay according to the intensity of care received each day (*WLOS*) and assigning to each NNU equally regardless of of the length or intensity of care *Equal*. The weightings are summarised for reference in table 7.1. Details for calculating the weights were provided in sections 5.3.2 and 5.4.2. Results are compared across weightings to check the sensitivity of inferences regarding NNU and network effects.

7.4 Results for analysis of neonatal data

7.4.1 Data

Table 7.3 shows the infants characteristics of the cohort analysed in this chapter. The data used in this chapter are a subset of the infants described in section 2.3.1. Consistent with the analysis of non-transferred infants presented in chapter 6 I excluded infants with missing risk adjustment data (n=217, see section 3.5 for a discussion of these exclusions). I also excluded 614 infants without a complete episodic record. This left a total of 15,538 infants in the analysis, 95% of the original cohort of 16,369. Altogether these infants had a total of 24980 records. Table 7.3 presents data separately for transferred and non-transferred infants as does the corresponding table prior to exclusion presented in table 2.2; results were similar before and after exclusions. Descriptive data were not presented by NNU level as the non-transferred cohort described in table 6.1 was because infants attend multiple NNUs across levels.

Table 7.3: Description of cohort for analysis of transferred infants by transfer status

		Non-transferred	Transferred	Total
Gestational age (weeks$^{+\text{days}}$)	$22^{+0} - 23^{+6}$	217(2.1)	144(2.7)	361(2.3)
	$24^{+0} - 25^{+6}$	697(6.9)	1006(18.5)	1703(11.0)
	$26^{+0} - 27^{+6}$	1375(13.6)	1330(24.5)	2705(17.4)
	$28^{+0} - 29^{+6}$	2767(27.4)	1458(26.9)	4225(27.2)
	$30^{+0} - 31^{+6}$	5054(50.0)	1490(27.5)	6544(42.1)
Birth weight (g)	< 500	68(0.7)	54(1.0)	122(0.8)
	$500 - 999$	2432(24.1)	2526(46.5)	4958(31.9)
	$1000 - 1499$	4554(45.0)	2015(37.1)	6569(42.3)
	$1500 - 1999$	2837(28.1)	763(14.1)	3600(23.2)
	$\geqslant 2000$	219(2.2)	70(1.3)	289(1.9)
Sex	Female	4618(45.7)	2384(43.9)	7002(45.1)
	Male	5492(54.3)	3044(56.1)	8536(54.9)
Steroids*	No	1256(12.4)	619(11.4)	1875(12.1)
	Yes	8854(87.6)	4809(88.6)	13663(87.9)
Death before discharge	Survived	9009(89.1)	4963(91.4)	13972(89.9)
	Died	1101(10.9)	465(8.6)	1566(10.1)

* Any dose of antenatal steroids given to the mother

7.4.2 Weight distribution

Table 7.4 compares how the total weight across all infants is allocated to the different NNU levels, shown for the whole cohort and by survival status. More care was attributed

to higher level NNU, similarly to the analysis in section 2.3.2. In particular, infants that died were mostly allocated to neonatal intensive care units (NICUs) (consistent with section 2.3.4) but the differences were less strong for the *First* weighting. Otherwise, deaths were distributed fairly similarly across NNU levels for weightings, though there was slightly less care allocated to NICUs under *Equal* weighting. The *WLOS* weighting allocated the lowest proportion of survivors to NICUs compared with other weights and, correspondingly, the highest proportion for special care baby units (SCBUs) and local neonatal units (LNUs). The *Beta3* weighting allocated the highest proportion of survivors to NICUs compared with other weights. There was no consistent pattern to the weights as more weight was placed at the beginning of stay (comparing *Beta2*, *Beta3* and *First*).

To show how the weight was distributed over the course of stay, table 7.5 compares how the total weight across all transferred infants is allocated to the first NNU, interim NNU(s) and last NNU they are treated in (infants only transferred once they will only have a first and last NNU). The *LOS* weighting places the most weight with the final NNU treating the infant. After the *First* weighting, the *Beta3* and *Beta2* weightings place the most weight on the first NNU, but the differences compared with the other weights are not large. Comparison of the *LOS* and *WLOS* weightings shows a difference for survivors, with the end of stay downweighted for the latter.

7.4.3 Model checking

Convergence Model details were described in section 5.4. All models were run with a burn-in of 20,000 iterations and a subsequent posterior sample of 200,000 iterations. Similar to the analysis presented in chapter 6, autocorrelation was high for some parameters so a large number of posterior samples were required to obtain the required effective sample size. Convergence assessment was described in section 5.4; the Monte Carlo standard error and effective sample size are provided in appendix F.

Model fit Table 7.6 shows the DIC, deviance and effective number of parameters for the difference MM models. There were substantial differences in DIC across models (minimum difference of 20), with the *LOS* weighting giving the lowest DIC and the *First* weighting the highest DIC. Here the models differ only in the NNU effects, so a higher effective number of parameters indicates larger difference between the NNU effects (i.e. a higher between-NNU variance, as will be discussed in section 7.4.5).

	SCBU	LNU	NICU
All infants			
Equal	9.4	40.0	50.7
LOS	9.4	40.2	50.4
Beta2	7.5	38.8	53.7
Beta3	6.5	38.1	55.4
WLOS	7.9	38.8	53.3
First	8.1	40.1	51.8
Survivors			
Equal	9.9	41.8	48.3
LOS	10.1	42.4	47.5
Beta2	8.0	40.9	51.0
Beta3	6.9	40.0	53.1
WLOS	8.5	40.9	50.6
First	8.2	41.0	50.8
Deaths			
Equal	4.5	23.8	71.6
LOS	3.0	20.4	76.6
Beta2	2.8	20.3	76.9
Beta3	3.0	20.8	76.2
WLOS	3.0	20.2	76.9
First	7.1	31.5	61.4

Table 7.4: Distribution of multiple membership weight across NNU levels

7.4.4 Infant and provider characteristics

Table 7.7 shows the posterior mean odds ratios (OR) and 95% credible intervals (95% CrI) for the infant characteristics. Results were fairly similar across all six weights, though different to those estimated for inborn infants alone in table 6.3. The odds ratios for male sex and antenatal steroids were both lower than that for non-transferred infants, but more similar to those in the single level model developed in chapter 3 (see table 3.5).

The ORs comparing mortality in SCBUs and NICUs with LNUs varied considerably across the six weights, with all results different from those for non-transferred infants, where both SCBUs and NICUs had higher mortality than LNUs. The *LOS* weighting gave the largest differences with the odds of mortality in SCBUs being half those in LNUs (OR 0.46 95% CI (0.28,0.75)), while the OR for NICUs was 2.68 (2.18,3.30). The *Equal* and *WLOS* weightings produced similar but less extreme results: *Equal* SCBU OR 0.62 (0.40,0.93) NICU 2.22 (1.82,2.73); *WLOS* SCBU 0.64 (0.38,1.04) NICU 2.01 (1.62,2.47). There was a pattern of decreasing mortality in NICUs as more weight was assigned to the early part of stay: *Beta2* 1.79 (1.48,2.18); *Beta3*: 1.33 (1.10,1.60); *First*:

	First	Interim	Last
All transferred infants			
Equal	40.3	19.4	40.3
LOS	26.6	18.7	54.6
Beta2	37.7	23.4	38.9
Beta3	45.1	25.3	29.6
WLOS	35.2	22.4	42.4
First	100.0	0.0	0.0
Survivors			
Equal	39.8	20.5	39.7
LOS	26.5	19.7	53.8
Beta2	38.2	24.7	37.1
Beta3	45.7	26.8	27.5
WLOS	35.8	23.7	40.4
First	100.0	0.0	0.0
Deaths			
Equal	45.9	8.0	46.1
LOS	28.6	8.1	63.3
Beta2	32.6	9.0	58.4
Beta3	38.2	9.1	52.7
WLOS	28.7	7.9	63.3
First	100.0	0.0	0.0

Table 7.5: Distribution of multiple membership weight across episodes, transferred infants only

	DIC	Deviance	Effective parameters
Equal	7644.00	7589.00	55.08
LOS	7572.00	7513.00	59.21
Beta2	7712.00	7663.00	48.72
Beta3	7758.00	7713.00	44.94
WLOS	7686.00	7633.00	53.06
First	7778.00	7736.00	42.23

Table 7.6: Model performance

0.93 (0.80,1.08).

Table 7.7: Posterior mean odds ratios and 95% credible intervals for infant and provider characteristics for multiple membership models

	Equal		*LOS*		*Beta2*		*Beta3*		*WLOS*		*First*	
	OR	95% CrI	OR	95% CrI	OR	95% CrI	OR	95% CrI	OR	95% CrI	OR	95% CrI
$GA^{(1)}$	0.15	(0.09,0.25)	0.18	(0.11,0.29)	0.17	(0.10,0.29)	0.15	(0.09,0.25)	0.18	(0.11,0.29)	0.15	(0.09,0.24)
$GA^{(2)}$	2.47	(1.60,3.78)	2.23	(1.42,3.47)	2.26	(1.43,3.55)	2.45	(1.58,3.74)	2.22	(1.43,3.41)	2.43	(1.61,3.74)
BWT	0.88	(0.71,1.10)	0.91	(0.72,1.13)	0.91	(0.72,1.14)	0.87	(0.70,1.09)	0.91	(0.73,1.14)	0.85	(0.68,1.06)
BWT^2	1.82	(1.62,2.04)	1.77	(1.58,1.99)	1.81	(1.61,2.03)	1.86	(1.66,2.08)	1.79	(1.60,2.01)	1.88	(1.68,2.10)
$GA^{(1)}\times$ BWT	0.37	(0.22,0.62)	0.40	(0.24,0.68)	0.40	(0.23,0.68)	0.36	(0.22,0.61)	0.40	(0.24,0.68)	0.36	(0.22,0.60)
$GA^{(1)}\times BWT^2$	0.80	(0.71,0.91)	0.82	(0.72,0.93)	0.81	(0.71,0.92)	0.80	(0.70,0.90)	0.81	(0.72,0.92)	0.80	(0.71,0.90)
BWT $\times$ $GA^{(2)}$	1.21	(0.84,1.74)	1.17	(0.80,1.69)	1.15	(0.79,1.68)	1.21	(0.84,1.73)	1.16	(0.80,1.67)	1.20	(0.84,1.73)
Male	1.43	(1.27,1.62)	1.42	(1.26,1.61)	1.42	(1.25,1.60)	1.42	(1.26,1.61)	1.42	(1.26,1.61)	1.42	(1.26,1.61)
Steroids	0.41	(0.35,0.48)	0.41	(0.35,0.48)	0.42	(0.36,0.49)	0.43	(0.36,0.50)	0.41	(0.35,0.48)	0.43	(0.37,0.51)
SCBU	0.62	(0.40,0.93)	0.46	(0.28,0.75)	0.73	(0.46,1.16)	1.14	(0.72,1.72)	0.64	(0.38,1.04)	1.09	(0.84,1.43)
NICU	2.22	(1.82,2.73)	2.68	(2.18,3.30)	1.79	(1.48,2.18)	1.33	(1.10,1.60)	2.01	(1.64,2.47)	0.93	(0.80,1.08)

OR Odds ratio; CrI Credible interval

$GA^{(1)}$ and $GA^{(2)}$ denote spline terms for gestational age at birth; BWT birth weight

All continuous coefficients were standardised so OR are for a one standard deviation increase

* SCBU odds ratio comparing special care baby units to local neonatal units (LNUs); NICU odds ratio comparing neonatal intensive care units to LNUs

7.4.5 NNU and network effects

Standard deviation of NNU and network effects

Figures 7.1 compares the between-NNU and between-network standard deviation (σ_u and σ_v) for each of the six weights. Figure 7.2 compares σ_u across all six weights and figure 7.3 does the same for σ_v. The posterior mean of σ_u was highest for the *LOS* weighting (0.31) and lowest for the *First* weighting (0.14). For comparison, the between-NNU standard deviation was 0.27(0.07,0.46) for the same model for non-transferred infants (without the MM, model 5 in chapter 6). For weights *Beta3* and *First*, which attribute the most weight to initial part of stay, σ_u was more uncertain with posterior mass near zero, see figures 7.1 (d) and (f).

The σ_v was fairly similar across models (posterior mean of around 0.23) but slightly less for the *First* weighting (0.20). This is similar to the result of σ_v 0.22 (0.03,0.4) in non-transferred infants. In contrast to non-transferred infants only (figure 6.6 (e)) σ_u was more uncertain than σ_v, and of lower magnitude for three weighting schemes (*Beta2*, *Beta3* and *First*), see figures 7.1 (c),(d) and (f).

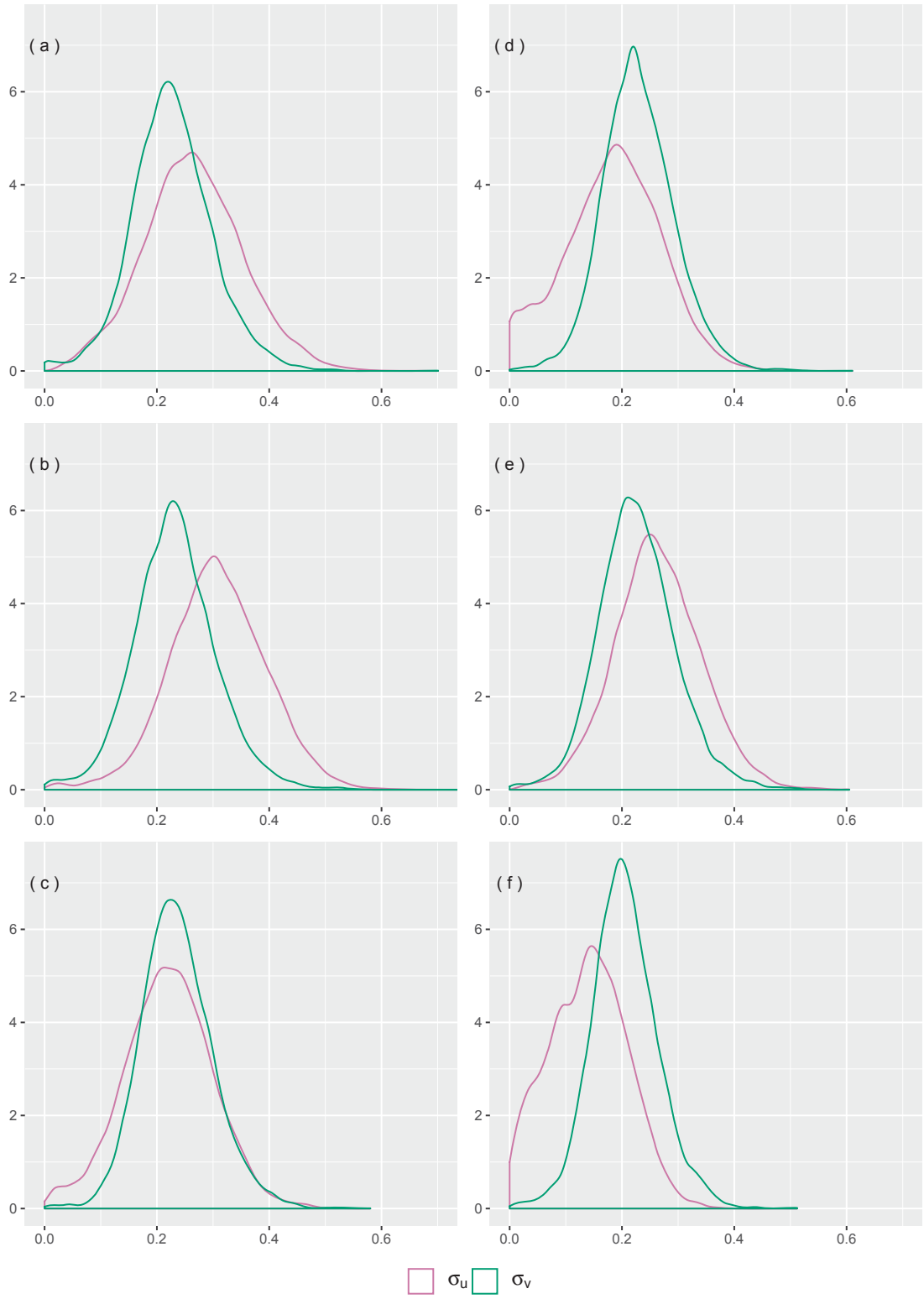


Figure 7.1: Posterior distribution for σ_u and σ_v for each MM weighting: (a) *Equal*, (b) *LOS*, (c) *Beta2*, (d) *Beta3*, (e) *WLOS* and (f) *First*

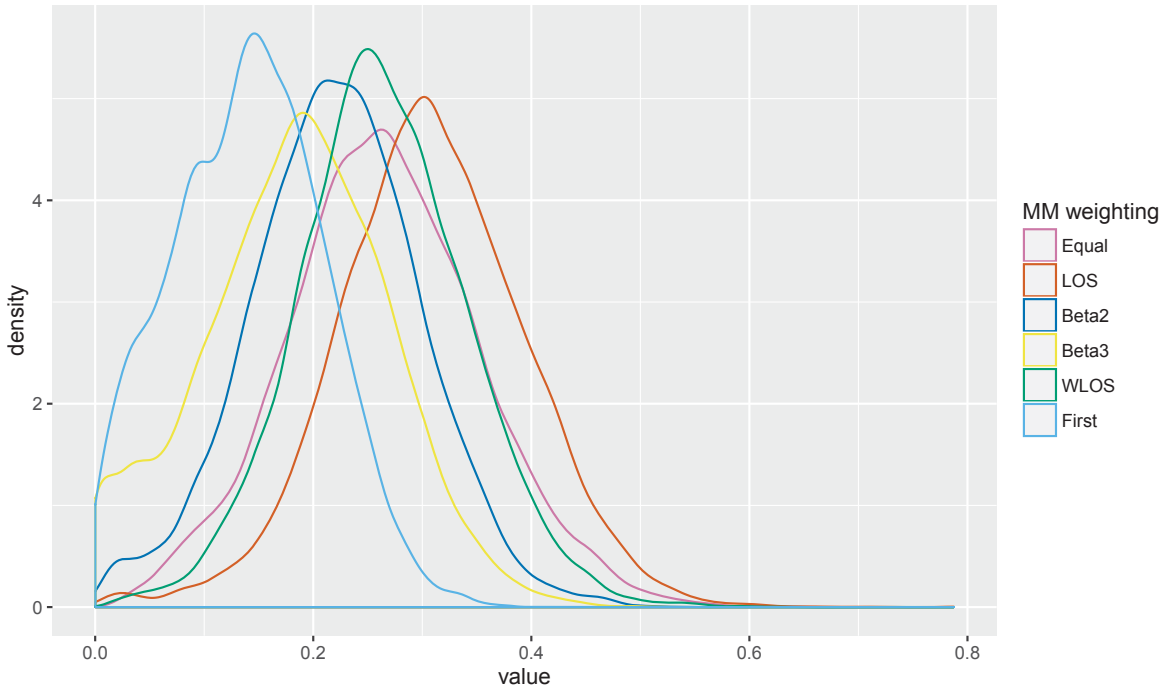


Figure 7.2: Posterior distribution comparing σ_u (between-NNU SD) across all multiple membership weights

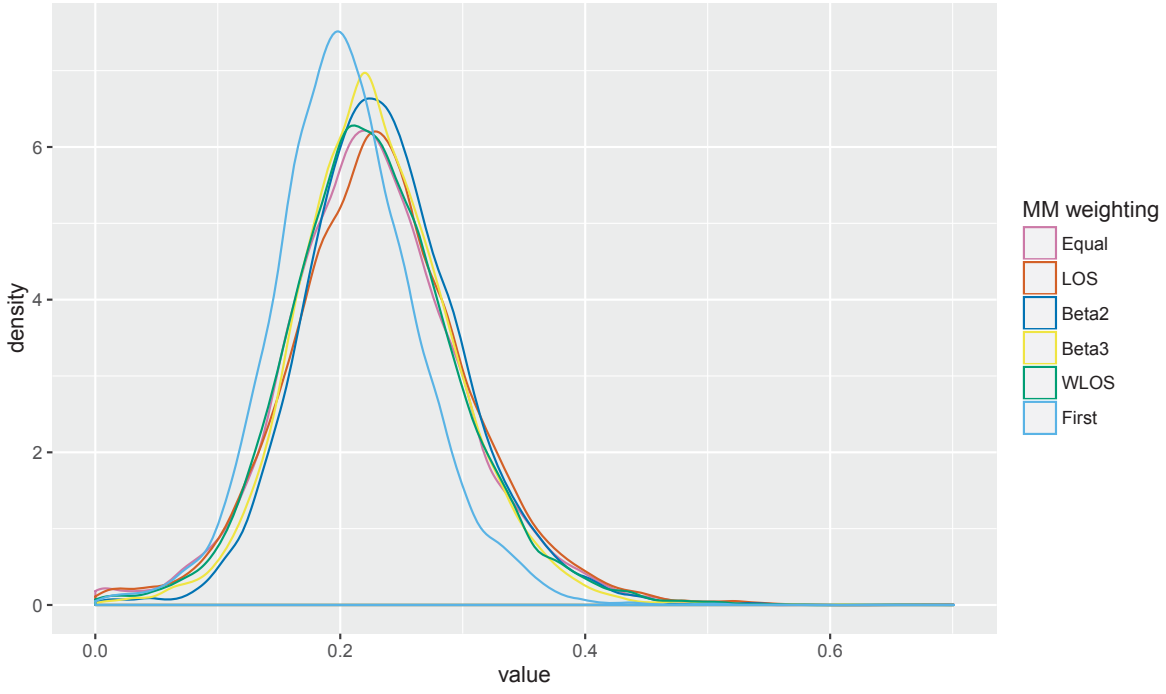


Figure 7.3: Posterior distribution comparing σ_v (between-network SD) across all multiple membership weights

Comparison of NNU effects across models

Figure 7.4 shows the correlation of individual NNU effects for each pair of weights, and the distribution of posterior means of NNU effects by NNU level. The lowest correlation (0.64) was between the *LOS* and *First* weightings and the highest correlation (0.98) was between the *LOS* and *WLOS* weightings. Overall the *First* weighting was not well correlated with the other weights (0.64-0.83) while the *WLOS* weighting correlated well with all weights (0.9 or greater) except *First*. Within NNU level, correlation between weighting schemes was generally lowest for SCBUs and highest for NICUs. Similarly to the analysis of non-transferred infants shown in figure 6.7, the lower level NNUs were more shrunk towards the mean. Shrinkage was most apparent in the *Beta2*, *Beta3* and *First* weightings, corresponding to the lower σ_u seen in these models. There were a few outlying NNUs which were particularly apparent for the *Equal*, *LOS* and *WLOS* weightings.

Table 7.8 shows the precision of the NNU effects, as measured by the average SD. The lower the σ_u (figure 7.2), the more information is borrowed and the greater the precision. NNU effects had the most uncertainty for SCBUs and the least for NICUs, consistent with differences in size.

Figures 7.5 to 7.8 show caterpillar plots for SCBUs, LNUs, NICUs and networks. Across all levels of NNU, the largest log OR were generally produced by the *LOS* weighting and the smallest by *First* and *Beta3*, consistent with the between-NNU SD (figure 7.2) and shrinkage of posterior means (figure 7.4). There was considerable overlap in the credible intervals, even for NICUs, with wider intervals for the larger effects produced by the *LOS* weight, consistent with the precision in table 7.8.

For most SBCUs there was little difference between the smallest and largest OR compared with the uncertainty. SCBU 36 had a substantial difference between the smallest and largest NNU effect, which can also be seen in figure 7.4. This NNU had 178 admissions, all of which were transfers in and all survived to discharge. Therefore the *Equal* and *LOS* weightings do not reflect the lower intensity of care, resulting in a very low (though uncertain) log OR for this NNU.

For LNUs the differences across models were slightly bigger than for SCBUs, but generally small relative to uncertainty. For NICUs there were more differences observed across the range with less of a consistent pattern. Credible intervals were a little smaller for NICUs, though still much larger than the differences across weights. There were a few LNUs and NICUs where both the largest and smallest estimates of the OR were on the same side of zero, but the credible intervals always included zero. As with the results in chapter 6, differences between weightings may be meaningful when expressed as the

probability of exceeding a certain threshold. For example for SCBU 104, the probability of the OR exceeding 1.2 (equating to a log OR of 0.18) is 31% for the *LOS* weighting but only 6% for the *First* weighting. Differences can also occur for NICU where the credible intervals are narrower; for example, for NICU 162 the probability of the OR exceeding 1.2 is 28% under the *Equal* weighting and *8%* under the *Beta3* weighting.

For LNUs and NICUs it was not always the case that NNUs with more extreme results had the bigger differences across weightings (as was observed when comapring models using caterpillar plots in chapter 6). This suggests that for a particular NNU, differences in the log OR across weightings are not solely driven by the overall between-NNU variance and amount of shrinkage, but also by how the weightings allocate care of infants treated in that NNU. This is consistent with the low correlation across weightings seen in figure 7.4 and suggests that choice of weightings will affect some NNUs differently to others. To investigate this I reordered the caterpillar plots by the proportion transferred, considering transfers in and out, and transfers within 24 hours. None of these plots showed any particular patterns. Figure 7.9 shows LNU effects ordered by the proportion of infants transferred out within 24 hours as an example, the remaining plots are provided in appendix F.

There were little differences between results for neonatal networks given the overall uncertainty, though the *WLOS* weight tended to produce effects closest to the null.

	All	SCBU	LNU	NICU
Equal	0.25	0.31	0.25	0.21
LOS	0.28	0.33	0.28	0.22
Beta2	0.22	0.25	0.22	0.18
Beta3	0.18	0.20	0.19	0.16
WLOS	0.25	0.30	0.25	0.20
First	0.14	0.15	0.14	0.13

Table 7.8: Posterior mean standard deviation for NNU effects, averaged over NNU, MM models

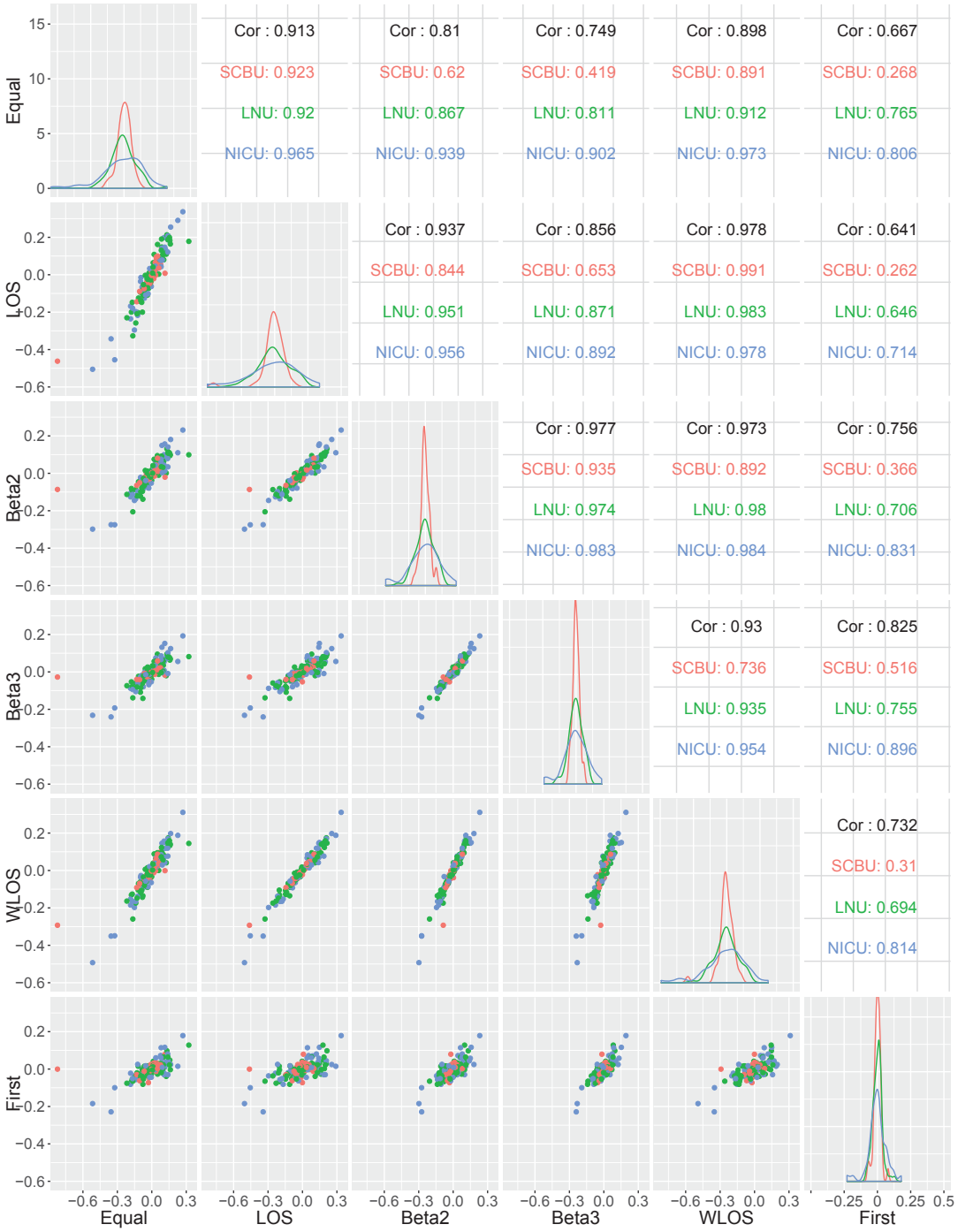


Figure 7.4: Correlation of NNU effects across multiple membership weights

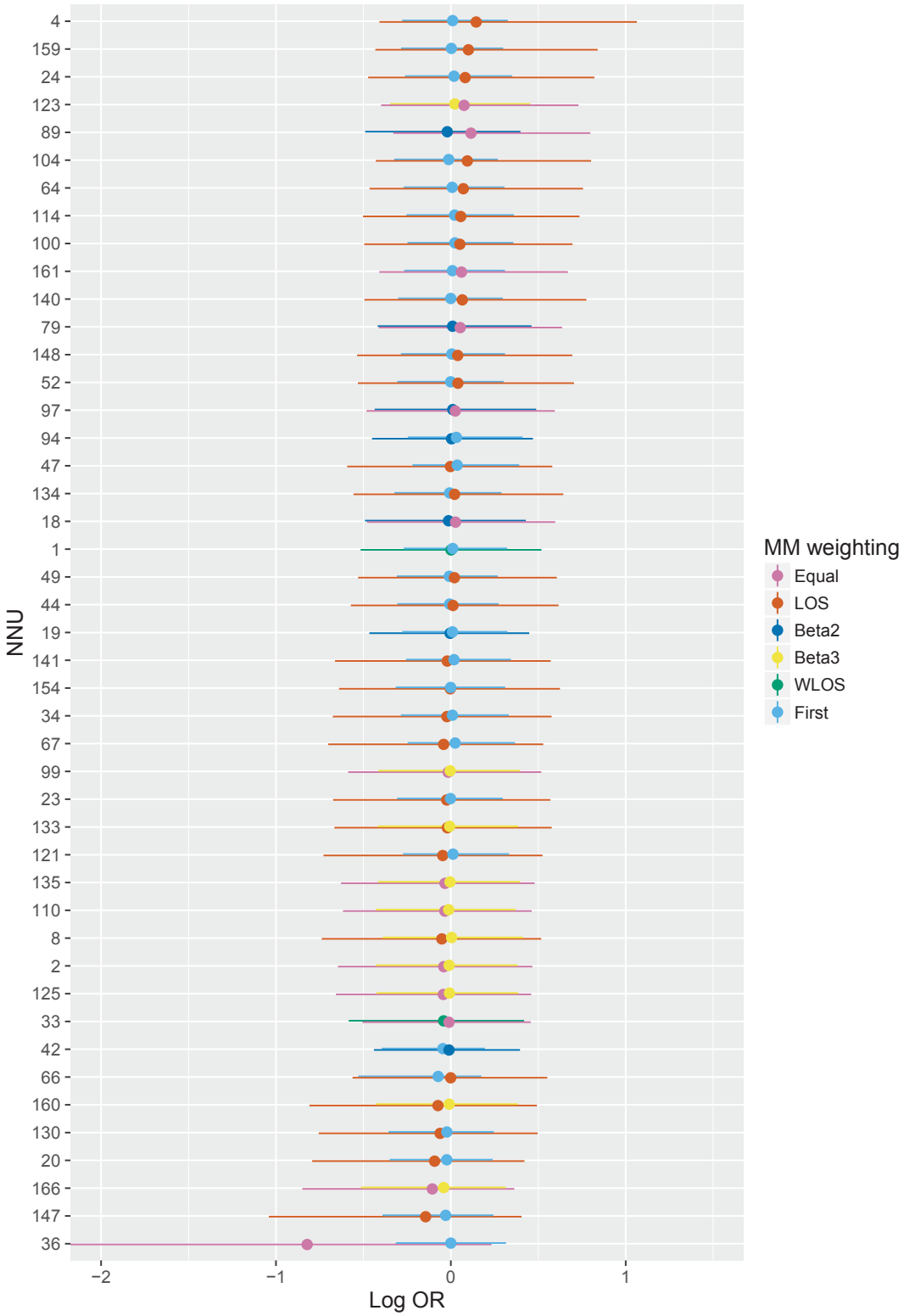


Figure 7.5: Caterpillar plot for SCBU effects showing smallest and largest effects across multiple membership weights

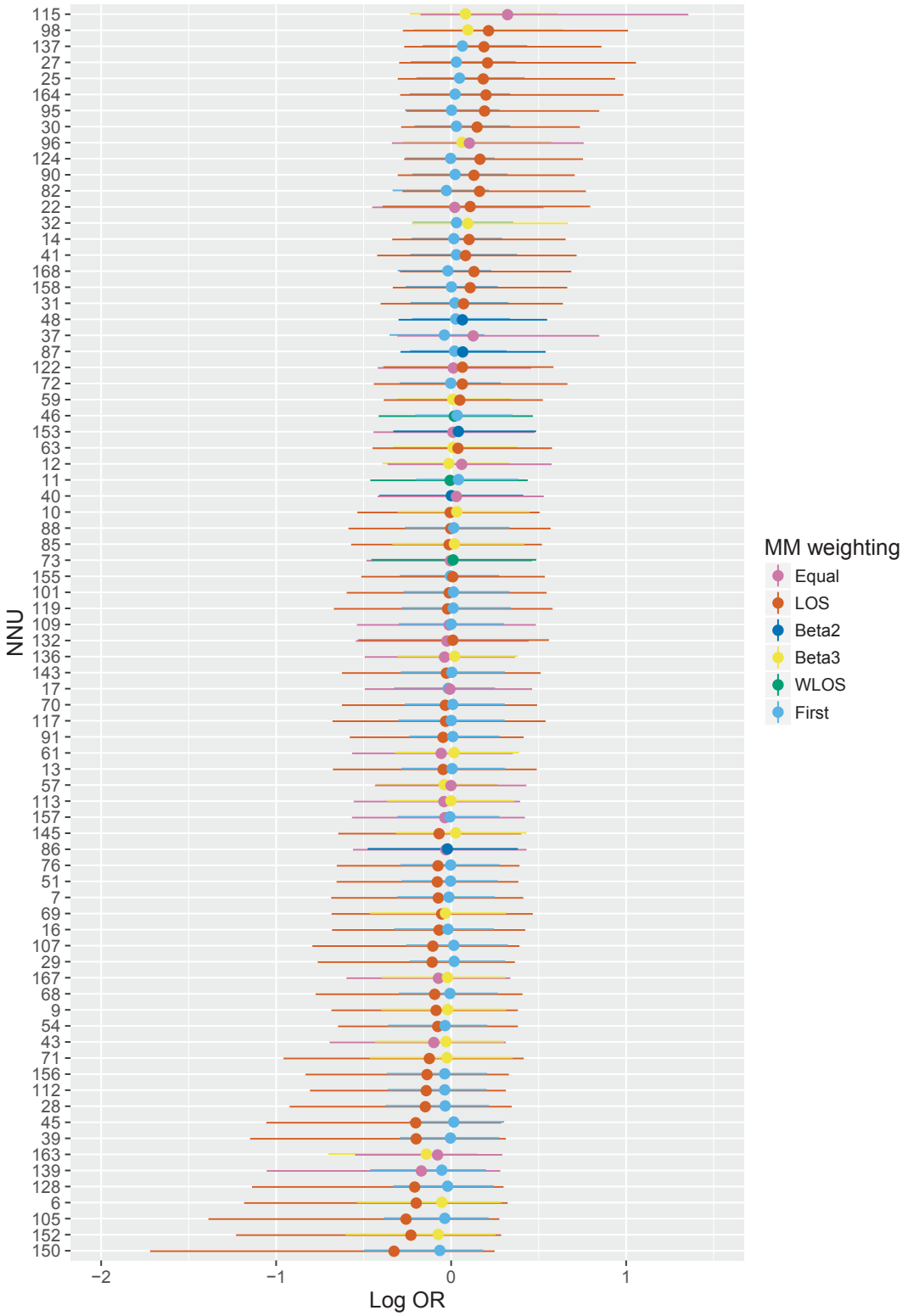


Figure 7.6: Caterpillar plot for LNU effects showing smallest and largest effects across multiple membership weights

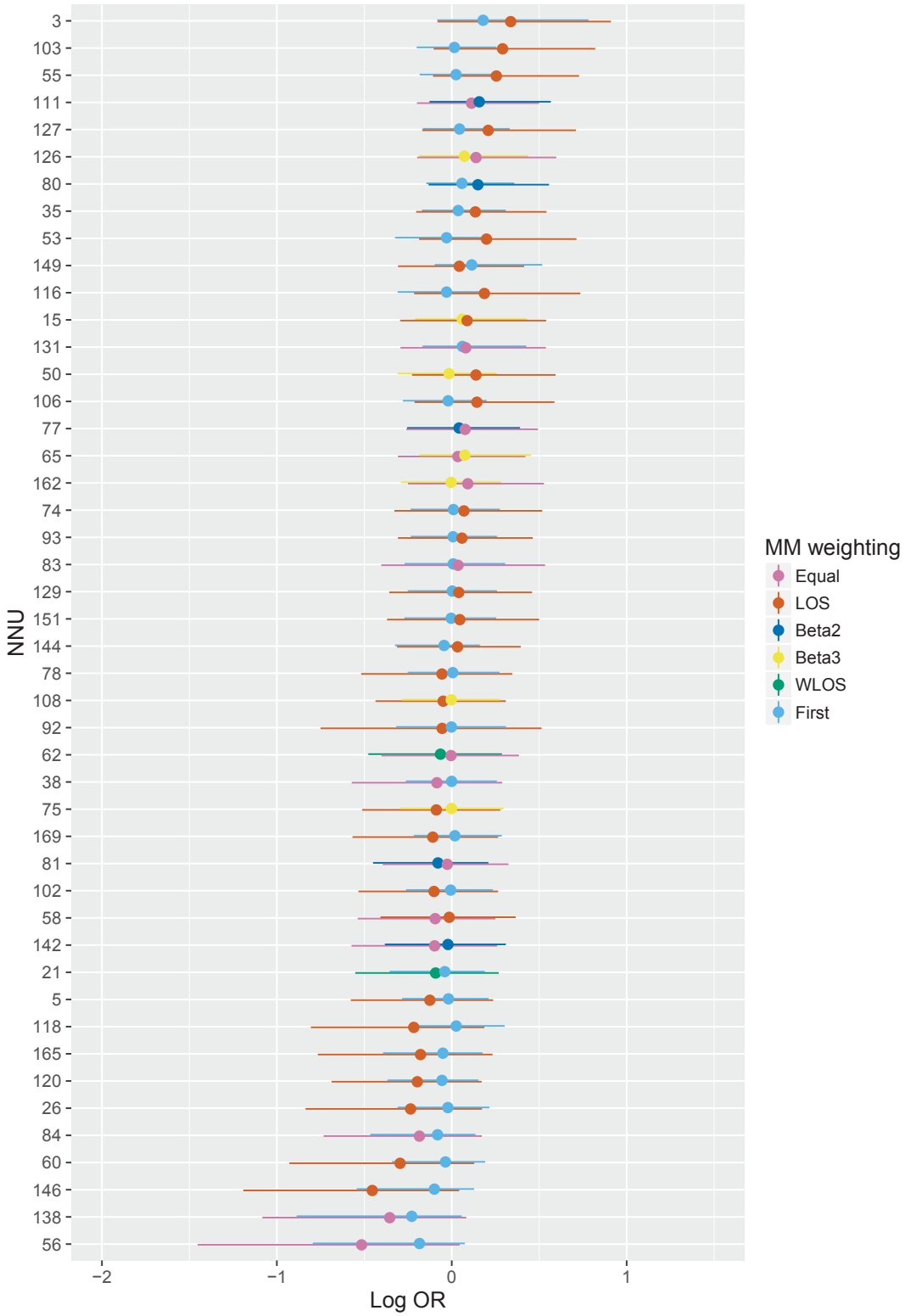


Figure 7.7: Caterpillar plot for NICU effects showing smallest and largest effects across multiple membership weights

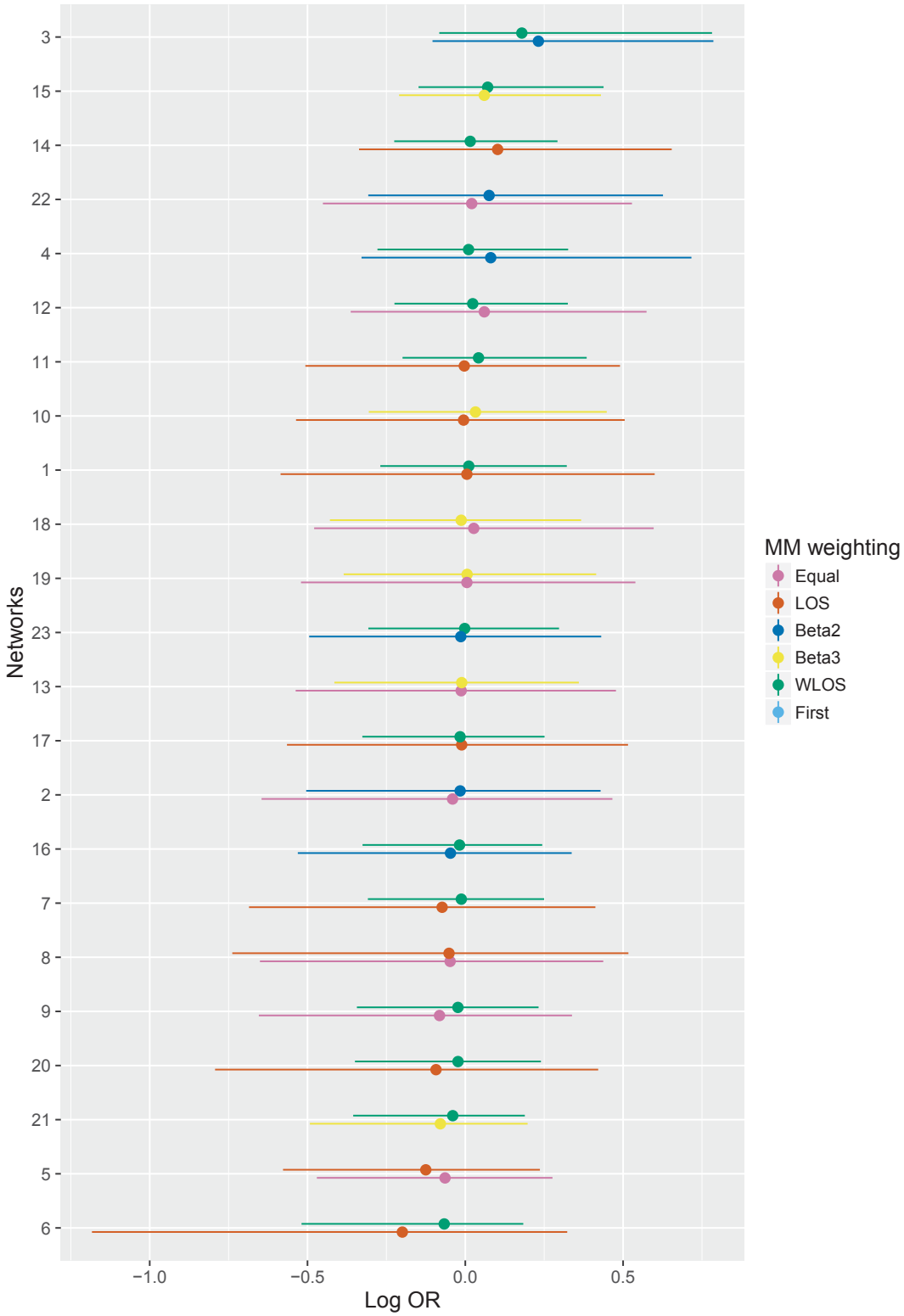


Figure 7.8: Caterpillar plot for networks effects showing smallest and largest effects across multiple membership weights

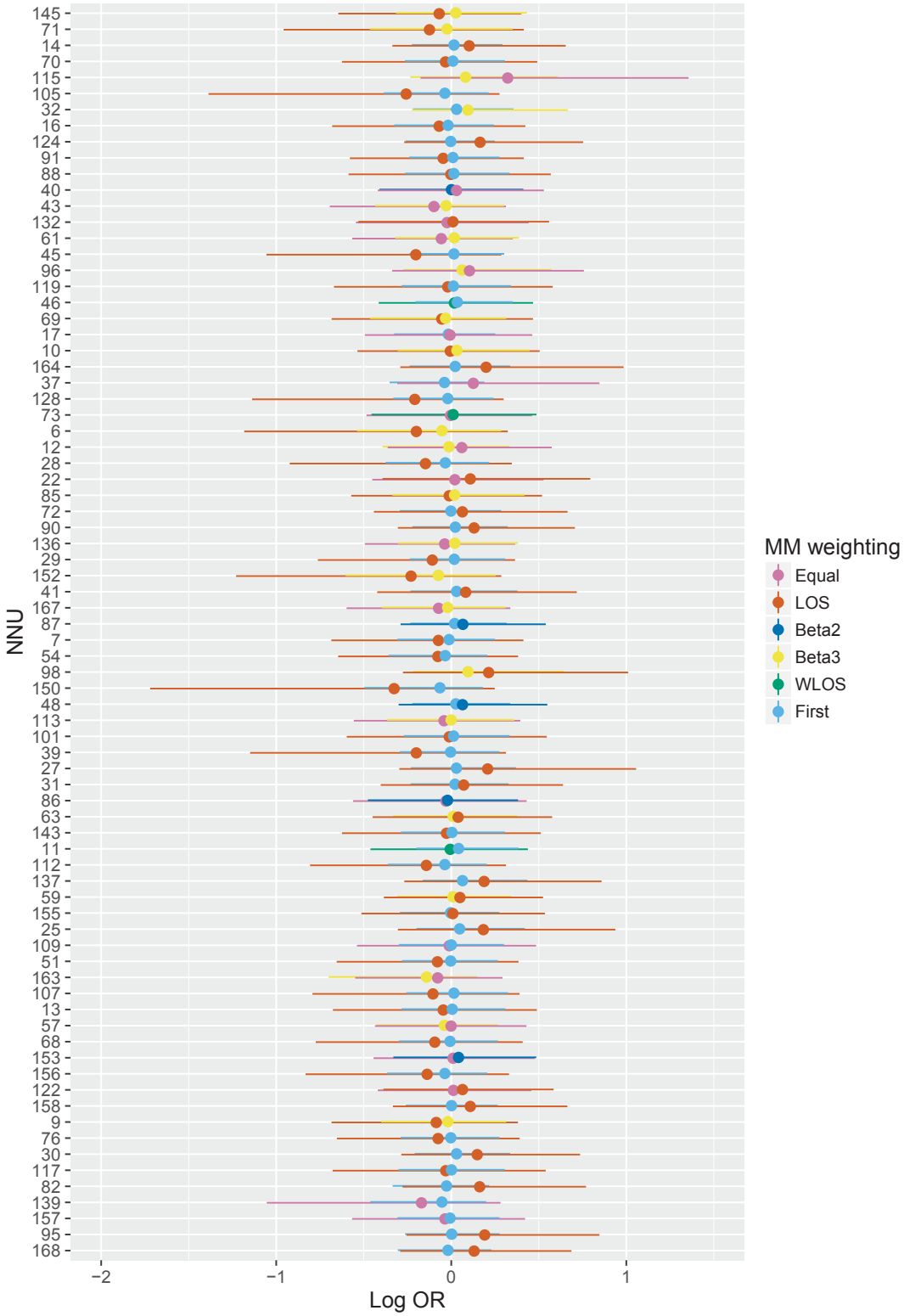


Figure 7.9: Caterpillar plot for LNU effects, ordered by the proportion of infants transferred out within 24 hours

7.5 Discussion

7.5.1 Summary

In this chapter I presented the results from the multiple membership analyses. First I presented the simulation study which showed that ignoring multiple membership structure and assigning individuals to a single provider resulted in underestimating the provider variance and poorer coverage of the variance and provider effects. Biases were larger when the weights generating the data were more evenly distributed than those used to analyse the data. In comparison, biases from assuming an incorrect MM weight were minimal. There was some indication of variance underestimation when a more uneven attribution was assumed but not the converse. Estimates of provider effects themselves were less strongly affected than the provider variance, with any biases overshadowed by shrinkage effects. While the average DIC was lowest for correct analysis models, the DIC had limited utility for model selection. In summary, multiple membership models were superior to allocation to a single provider, even when the analysis weighting differed to the scenario weighting, and individual provider results were not strongly affected by different weights.

I then applied multiple membership models to the neonatal data, extending the analysis presented in chapter 6. Associations with infant characteristics were slightly different from those found for the analysis of transferred infants only, though the risk profile and crude mortality was also different. In contrast with the analysis of non-transferred infants, mortality was in general highest for NICUs and lowest for SCBUs. This was consistent with the descriptive analysis showing that deaths tended to occur in NICUs and discharges home tended to occur from SCBUs after convalescent care. The ORs for NNU level were sensitive to the MM weights with the largest differences for the weight proportional to LOS.

As was seen in the simulation study, allocating infants to the first NNU only resulted in the lowest estimate of the between-NNU SD (σ_u). The standard approach of assigning infants proportional to length of stay in the NNU yielded the highest estimate of σ_u, and similar values were obtained using weighted LOS and equal allocation. The between-network variation was fairly similar across MM weights. The DIC was lowest for the weight proportional to LOS with substantial differences between models, much greater than those seen in the simulated data.

The NNU effects arising from allocation to the first provider did not correlate well with those from other MM weights and were shrunk more heavily towards the mean (as were those from the other weights placing more emphasis at the start of stay). The increase shrinkage meant that more information is borrowed and the uncertainty around

the NNU effects is lower. Hence the NNU effects from the first only weight produced posterior mean NNU effects which were smallest in magnitude, while the largest were generally produced by the LOS weight. Differences across weights were small for the most part and were overshadowed by the uncertainty. There were no clear patterns when examining the NNU effects by the level of transfers.

In summary the different aspects of the analysis produced a consistent pattern across the weighting schemes. The standard MM weighting of allocating in proportion to LOS produced the largest differences between NNU levels and the largest NNU effects. From a modelling perspective, this weighting also gave the lowest DIC and the largest between-NNU variance, and placed the most weight on the discharging NNU. In contrast, the standard provider profiling approach of allocating all infants to the first NNU produced opposite effects: the smallest differences between NNU levels and NNU effects, the largest DIC and the smallest between-NNU variance.

7.5.2 Simulation study

In line with previous MM studies (Chung and Beretvas, 2012; Leckie, 2009), the between-cluster variance was underestimated when a conventional hierarchical model was applied, assigning individuals to the first provider and thereby ignoring the MM structure of the data. The bias was much smaller when MM models were used, but notably the between-cluster variance was underestimated when the allocation of weight was more uneven than was implied by the model used to generate the data. The converse was not the case (use of a more uneven weight did not result in overestimating the variance). These results are drawn into question because using analysing the data using the MM weights that generated the data did not consistently minimise the bias or MSE for σ_u, which we would expect. The bias was calculated by comparing $\hat{\sigma}_u$ with the value of 0.4 used to generate the data, but random variation when sampling the hospital effects may mean that $\bar{\hat{\sigma}}_u$ is higher than 0.4, particularly as only 200 datasets were used. In this case any MM weighting which underestimates the between-cluster standard deviation will result in a lower bias. This could be further investigated using a higher number of simulations, which would reduce sampling variation and bring $\bar{\hat{\sigma}}_u$ closer to 0.4. In support of the findings, Wolff-Smith and Beretvas (2014) also found that a more uneven weight did not result in overestimating the between-cluster variance (see section 4.3.2), though the authors did not comment on why these results might occur.

Allocation to a single provider could be seen as an extreme case of uneven weighting. Returning to equation 4.7, the variance contribution under MM is scaled by the sum of the squared weights which is minimised when the weights are equal. This also implies that misspecification of the weight would lead to overestimation of the variance if the

applied MM weight is more even than than the underlying process generating the data. This was not seen in the simulation study here, and while some overestimation was observed in the study of Wolff-Smith and Beretvas (2014), in those scenarios application of the "correct" weight also yielded too high a variance.

Estimation of the provider effects themselves were slightly less biased with smaller mean square error when the correct analysis model is chosen, but differences are small and coverage across all MM models was good. Biases were larger and coverage was poorer in the simulated data when the MM structure was ignored. This has not been demonstrated in previous studies of MM models, but is important for provider profiling. Goldstein et al. (2006) used the DIC to select the weighting scheme, though I found that comparing DIC did not reliably distinguish the weight used to generate the data.

For provider profiling purposes we are not always interested in estimating the exact process generating the data. For example we may want all providers to be equally accountable, even though this may not reflect the influence each provider has on the outcome. This seems a reasonable strategy to take, as results from the simulation study showed that provided MM models were used, the differences between them were minimal regardless of the true process. The results also suggest that allocating the weight equally amongst providers is the least likely to results in biased variance estimates so is a suitable default choice in the absence of a *a priori* weight distribution.

The conclusions drawn above may not be generalisable outside of the simulation study. The study was limited as I only applied a narrow range of assumed true weights and analysis models, and did not vary other simulation parameters. In particular the influence of the size of the dataset (number of individuals and number of providers), the proportion transferred, the number of times individuals were transferred and the between hospital variance were kept constant, and patient volume was equal across hospitals. This has the advantage of being easier to interpret, but limits the utility of the conclusions which can be drawn. To extend the simulation I would consider using the real neonatal data but simulating the NNU effects to more accurately reflect transfer patterns.

7.5.3 Analysis of neonatal data

For the neonatal data there were notable differences in the variance estimates and DIC across the different MM weightings, as well as differences in direction of NNU level effects. This is in part due to differences in transfer patterns across infants. Whether an infant is transferred, when transfer occurs and where the infant is transferred to are all influenced by health status before birth, at birth and beyond. Transfer patterns were

apparent from findings earlier in the thesis: in chapter 2 I demonstrated that infants who require a higher level of neonatal care than is available at their hospital of birth are transferred early, usually within 24 hours of birth. Infants whose condition improves are transferred later on for step-down care and are discharged from lower level NNUs, while infants that die tend to do so in NICUs. The aim of estimating NNU effects is to compare mortality at an NNU with what would be expected for similar infants at similar NNUs, but we cannot infer that any variation is caused by the NNU itself as there may be unmodelled confounders contributing to the NNU effects. The transfer patterns contribute additional between-NNU variation which is not caused by the NNU, but occurs in the opposite direction; that is the infant's condition influences which NNU they are treated in rather than care provided by the NNU affecting the infant's condition.

The standard MM weighting of allocating in proportion to LOS (here denoted *LOS*)showed the largest variation between NNUs and the largest differences between NNU levels. This weighting places the greatest weight on the discharging NNU (table 7.5) and thus leads to greater separation as to which NNUs survivors and deaths are allocated to. Table 7.5 showed the *WLOS* weighting mitigates this by downweighting the lower-level convalescent care for survivors. As infants that die are unlikely to have many days that are not intensive care, the difference in the *LOS* and *WLOS* weightings are negligible for them. The *Beta2, Beta3* and *First* weightings place much less weight on the discharging NNU, and as a result show the least variation between NNUs and the smallest differences between NNU levels. The *Equal* weighting to some extent avoids this problem: if an infant is born at SCBU A and transferred to NICU B for intensive care, then the equal weight assignment is identical whether or not the infant survives to return to SCBU A for convalescent care. Other selection effects remain: if the infant had not survived long enough to be transferred then all the weight would be allocated to SCBU A, but in such cases all weights would be the same as the infant is only treated in one NNU.

Some of the findings warrant further investigation. NNU effects for SCBU are more susceptible to differences in weighting (as shown by the lower correlation in figure 7.4). As shown in chapter 2 a higher proportion of SCBU admissions are transferred, but differences would be stronger if there is variation in the success of stabilising infants, as suggested by the higher (but very uncertain) between-SCBU variation discussed in 6.6. Despite allocation of weight for transferred patients being the only difference across models in this chapter, comparison of results for individual NNU by level of transfer using a variety of measures did not show any patterns. These are blunt summaries of complex patient pathways; a more detailed analysis, perhaps for the NNUs in one network, might help elucidate the reasons for difference in results across MM weights.

Comparison with previous findings

In comparison with the results for the same model for non-transferred infants (model 5 in chapter 6), the ORs for the infant characteristics were fairly similar, but the ORs for NNU level were very different. The between-NNU variation was slightly lower here for most models. There are a number of possible for this: differences between the MM weightings used to analyse the data and the underlying patterns of variation in the data (as discussed in section 4.3.2); differences in the population; or genuine difference in effect for transferred and non-transferred infants. Analysis of transferred infants only may help in understanding this, but this might not be feasible as the sample size would be reduced to a third.

In the simulation study differences between MM weightings were minimal regardless of the underlying process, but this was not the case for the neonatal data. There are a number of differences between the simulated and neonatal data: the variation attributable to the reverse causal effect may be large in comparison with NNU effect on mortality; the patterns generating the data are heterogenous across infants; and those patterns are very different to the analysis weights used. Differences in the effects of NNU level were seen between MM models but comparisons with simulated data could not be made as there were no provider-level variables in the simulation study. A more elaborate simulation study using the NNRD data including transfer patterns but controlling the NNU effects could be used to investigate this, as was discussed in 7.5.2.

Substantial differences across weighting schemes were not seen in most MM studies looking at pupil mobility (Fielding and Yang, 2005; Goldstein et al., 2007; Leckie, 2009). The destination school of a transferring pupil is perhaps more likely to be due to external reasons (e.g. moving house) than educational attainment. In contrast, Wolff-Smith and Beretvas (2014) find differences in the between-school variance when different weights are applied and suggest the possibility that pupil mobility could be affected by educational attainment, though the mechanisms for this are not discussed.

7.5.4 Conclusion

Hierarchical models can be put to different purposes. In many cases, including some of the existing MM studies previously described, we want the random effects to capture the variation in the data and choose models that appear to fit the data best. For provider profiling purposes we want the random effects to reflect the inferences we want to make about the provider effects, as was discussed in chapter 4. This is quite a different aim, and, as shown in this chapter, may not correspond to the models selected based on measures of fit such as the DIC. Models which capture the most variation may not

exclusively model the source of variation that we are interested in, and therefore could yield misleading conclusions.

The eventual purpose of these results is to inform NNUs how the expected mortality for infants in their care compares with that of similar infants attending similar NNUs. As with all provider profiling studies, any differences could be due to performance but they be also be due to unmodelled factors, in this case transfers. For transferred infants, we cannot consider the infants at different NNUs to be *similar* because the location of care is determined partly by their condition in a way that was not known at birth. This issue is characteristic of any networked system of care, as the transferred patients at larger centres will primarily be the sickest patients from outside the locality, and therefore will either die at the larger centre or survive to be transferred back when their condition improves. Given the importance of monitoring clinical outcomes and the increasing use of centralisation it is necessary to clarify what these methods can achieve and what the caveats are. Alternative methodological approaches will be briefly considered in 9.4.3.

Equally attributing the weight across all NNUs caring for an infant regardless of length of stay, timing of stay or intensity of care gives an intuitive interpretation of assigning equal accountability while reducing the extent to which the reversal causal association is captured. For neonatal care, the weighted LOS approach combining care level data and recommended staffing ratios provides one suitable comparison as the resulting NNU effects will be more influenced by infants to whom more care was provided. Results should be considered alongside those for non-transferred infants only as these provide NNU effects which are not affected by postnatal transfer patterns. Other data on case-mix and volume of care may also help interpret results. An example of how these results might be presented graphically for a single unit is given in chapter 8.

More generally, equal weighting is a reasonable default choice given the findings of the simulation study. Equal weighting is simple to implement and explain to non-statistical users and do not require any arbitrary numerical choices or thresholds. Sensitivity analysis using a range of weights to reflect both desirable interpretations of provider effects and plausible clustering patterns in the data is recommended. This is particularly important where the transfer pattern may be related to the provider effects themselves, the outcome, or both, as is the case here. Understanding the drivers for the transfer patterns is crucial to interpreting any differences between weights which may arise.

In this chapter I presented results from multiple membership models using a variety of weights applied to simulated and neonatal data, and discussed how these could be implemented in practice for provider profiling. Results from the equal and weighted LOS models in this chapter are presented for a single NNU as part of a case study in chapter 8 alongside results for non-transferred infants and other descriptive analysis.

The findings from this chapter will be discussed in the context of the whole thesis in chapter 9.

Chapter 8

Case study

8.1 Aim

This chapter is a case study to show how the results arising from chapters 6 and 7 might be presented graphically alongside other descriptive data from chapter 2 for a single neonatal unit (NNU) in a comparable way.

8.2 Methods

The overall cohort for this chapter was the same as that described in chapter 2: infants born 2011-2013 at or below 31^{+6} weeks gestation and admitted to an NNU in England contributing to the National Neonatal Research Database (NNRD). As for previous analyses, infants with a birth weight greater than 4 standard deviations from the gestation and sex specific mean (based on the Royal College of Paediatrics and Child Health preterm standards (Wright et al., 2010) were excluded as were those with missing death data. For descriptive analyses every unique admission was counted, i.e. transferred infants were assigned to all NNUs caring for them, but readmissions of the same infant to the same NNUs were not counted twice unless otherwise stated. Non-transferred infants were assigned to the sole NNU providing care. For each analysis, infants with missing data for the variable in questions were excluded. The results for the NNU of interest (NNU X in neonatal network Y) are shown in the context of all other NNUs, NNUs at the same level (in this case Neonatal Intensive Care Unit (NICU), the highest level, see 1.2 for definitions) and, for volume data, NNUs in the same network. The NNU of interest is highlighted in blue and all others are shown in grey.

8.2.1 Volume

Volume of care was calculated in three ways: the number of infants in the cohort who were admitted to each NNU over the three-year period; the number of care days of any type provided by each NNU (where a care day is one day of care provided to one infant); and the number of intensive care days provided (the highest level of care, see 1.2 for definitions). For these plots the NNUs are displayed on order of the measure of volume, with NNU X highlighted. For each admission to NNU X (including readmissions of the same infant) I described the source of the admission: whether the infant was inborn (the admission to NNU X was the first admission to neonatal care), transferred from one of the six other NNUs in the same neonatal network (two special care baby units (SCBUs), three local neonatal units (LNUs) and one NICU), or transferred from an NNU in a different network.

8.2.2 Population

A selection of population attributes were compared between NNU X and the overall patient population, firstly for all NNUs and for NICUs only. Six factors were described: gestational age, birth weight, estimated risk of death, the level of socioeconomic deprivation in the maternal area of residence, maternal ethnicity and maternal age. Of these, the last three were not adjusted for in the mortality analysis described in section 8.2.3. Estimated risk of death was calculated using the logistic regression model described in chapter 3. For socioeconomic deprivation, the Index of Multiple Deprivation (IMD) of the Lower Super Output Area (LSOA) is used, which is an area-level composite measure, see section 2.3.1 for more details. The LSOA is determined from postcode of the infant's mother, and all LSOAs in England are ranked according to the IMD score. The decile of the rank is the measure used in the analysis. Gestational age, socioeconomic deprivation and maternal ethnicity were described using histograms and birth weight, risk of death and maternal age using density plots.

8.2.3 Mortality

Mortality data was shown separately for non-transferred infants, and for all infants together reflecting the separate analyses of chapters 6 and 7. The unadjusted mortality across all NNUs for non-transferred infants (i.e. those whose sole admission was to NNU X), was calculated as the number that died over the total. For all infants, this is calculated including deaths in the numerator for all NNUs treating the infant, not just the NNU in which the death occurred.

A risk-adjusted measure of mortality *for non-transferred infants* in each NNU is given by the odds ratio (OR) calculated using model 5 in chapter 6, full details of which can be found therein and statistical methods in section 5.2. The model used is a three-level hierarchical Bayesian logistic regression model (infants within NNUs within networks), adjusted for gestational age, birth weight, sex, whether the mother was given any antenatal steroids and level of NNU. The adjusted OR for NNU X can be interpreted as how the expected odds mortality for infants in their care compares with that of similar infants attending similar NNUs, with values above 1 indicating higher mortality than expected. The estimate we obtain is the mean of the posterior distribution, which describes the probability distribution for the OR. The 95% credible interval describes the 2.5% to 97.5% range of the distribution; this is provided for NNU X as an error bar to give a measure of uncertainty. Using a hierarchical model means that the ORs are more reliably and precisely estimated as they are shrunk towards the expected value for that provider. It also means clustering in the data is properly accounted for. All infants are non-transferred so they can be assigned to one NNU.

A risk-adjusted measure of mortality *for all infants* in each NNU, including those who are transferred, is given by the OR calculated using the multiple membership (MM) models in chapter 7. Like the model used to determine the OR for non-transferred infants, these are three-level models adjusted for the same infant and provider characteristics. In contrast, rather than just being assigned to one NNU infants are assigned to all NNUs that treat them using weights. Two models are presented here corresponding to different MM weights and hence different interpretations. The first weighting is equal across all NNUs providing care, regardless of what type of care was provided and for how long. This approach for assigning is similar to that for unadjusted mortality. For the second weighting infants are assigned relative to the duration and intensity of care at each NNU so the resulting OR gives more responsibility to NNUs providing a greater share of care. These weightings are denoted *Equal* and *WLOS* respectively as was described in chapter 7; more details on the MM models and calculation of the weights can be found in section 5.4.

I have briefly summarised the results presented in the graphs but clinical inferences for the NNU are minimal as the case study is primarily for illustrative purposes.

8.3 Results

The results are shown in figures 8.1 to 8.33.

NNU X is one of the larger NNUs in terms of volume, particularly with regard to the amount of intensive care conducted. NNU X is also one of the larger NICUs in relation

to intensive care volume and conducts much more intensive care than all the other NNUs in network Y (figures 8.1 to 8.9). More infants are transferred in to NNU X than are inborn, and more infants are transferred out for further care than are discharged home (figures 8.10 and 8.11). The gestational age distribution for NNU X is quite different to that both overall and for other NICUs (figures 8.12 and 8.13). The distribution appears bimodal, with a peak at 26 weeks gestation and another at 30 weeks gestation. Infants admitted to NNU X have lower birth weight than NNUs overall but a similar distribution to the NICU population (figures 8.14 and 8.15). In contrast infants in NNU X are more vulnerable based on estimated risk of death when compared with other NICUs, as well as other NNUs in general (figures 8.16 and 8.17). Compared with other NNUs and NICUs, the mothers of infants treated at NNU X are less likely to come from the most or least deprived areas , less likely to be of white ethnicity, and are slightly older.

Mortality for NNU X is high compared with all NNUs but less so for NICUs (22% for non-transferred infants and 13% for all infants) before using regression models to adjust for risk factors and shrink the NNU effects. This changes after adjustment, with all mean posterior odds ratios for NNU X being less than one. While there are some differences in the adjusted ORs for NNU X between the three adjusted models (for non-transferred infants, for all infants using the *Equal* MM model and for all infants using the *WLOS* MM model), the estimates are all very uncertain and span much of the range of NNU effects. The upper limit of the credible interval is lower for for both MM models compared with the model for non-transferred infants, particularly for the *WLOS* model. This means that probability of NNU X having a high odds ratio (greater than 1.2, say) is lower when all infants are included, particularly when we consider the amount and level of care provided.

8.4 Discussion

In this chapter I combined results from the thesis for a single NNU to give an example of how they might be presented graphically for provider profiling purposes. Clinicians could use this information locally to understand whether the NNU mortality rate is higher or lower than expected, the strength of the evidence behind this and how this is relates to information about the NNU, such as deprivation and volume. A commissioner might look at these results to see if differences in clinical need and outcomes across the NNUs under his or her remit matches funding and resource use. These results are not be suitable for the purposes of patient choice as, for reasons described in 4.4.2, they should not be used for direct comparison of individual providers.

The manner of presentation is similar to that used by the Health and Social Care Information Centre illustrated in figure 3.1. I have only presented the figures here, but

these would need to be combined with a summary of the data, methods, and advice on interpretation (including caveats) to form a coherent report. The nature of this and the level of detail would depend on the audience. The data presented could be extended (other clinical data for medical staff, other transfers data for managers), as could the weights used in the MM models. For example if staff at a SCBU are interested in outcomes for infants they stabilise, all the weight (or an increased proportion) could be placed on the first day. Similar reports could be constructed for a network, both for NNU effects with all NNUs in the network highlighted, and for network effects.

In all cases the adjusted ORs are the exponentiated NNU effects, which compare the NNU with an average NNU, adjusting for infant and provider characteristics. This may not be the desired comparison - for example we may have reason to compare across all NNUs regardless of level, and so do not want to adjust for this. To do this we can exponentiate the sum of the NNU effect and relevant NNU level regression coefficient arising from the hierarchical model to obtain the relevant OR. While this has not been implemented here, it shows the flexibility of the modelling approach to make different comparisons across providers.

There are drawbacks to presenting the ORs for mortality in this way. These plots rank the data, and estimated ranks are known to be unstable. Furthermore, ranking encourages direct comparison between NNUs which is not a reasonable inference for the methods used here (see 4.4.2). For these reasons I did not name the other NNUs, give explicit ranks, or present mortality data for the NNUs in the same network only (as there are only seven). I added the 95% credible interval for the OR to show the high levels of uncertainty. If results are produced separately and confidentially for each NNU and not in a single report, then the focus is on comparing a set of data for one NNU rather than one outcome across many NNUs, and conducting direct comparison would be difficult. Alternatively, the Bayesian methods applied in this thesis allow presentation of probabilities of interest directly, along with the associated uncertainty, which can be easier to interpret by users than frequentist measures (Normand et al., 1997). For example we could present the posterior probability of being in the top quantile of providers, the posterior probability of having an OR (exponentiated NNU effect) greater or less than a threshold, with input from clinical and managerial staff on the appropriate quantile/threshold to apply. If ranks were of interest these could be provided with credible intervals to show the uncertainty (Goldstein and Spiegelhalter, 1996). Despite the limitations and the relative complexity of the underlying models, the case study provides a clear picture of where an NNU sits in terms of its workload, population and mortality data, with a common method of presentation allowing these aspects to be assessed simultaneously.

Fig. 8.1 Number of infants admitted to NNU X (blue) and other NNUs (grey)

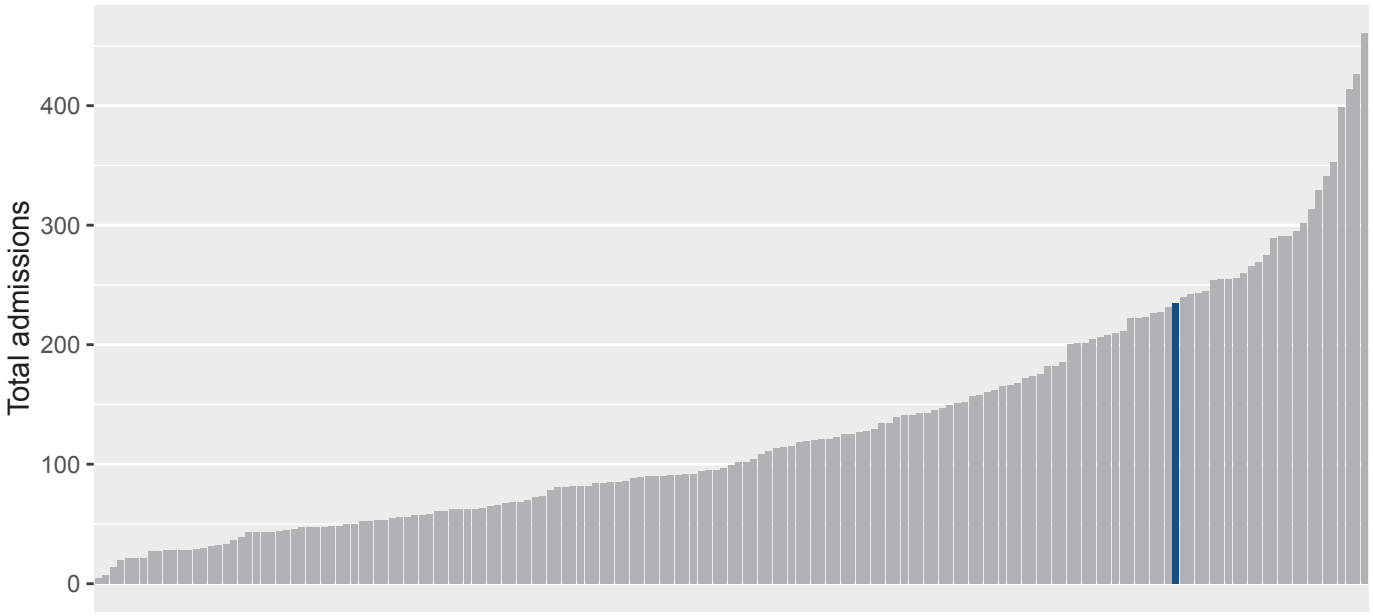


Fig. 8.2 Number of care days for NNU X (blue) and other NNUs (grey)

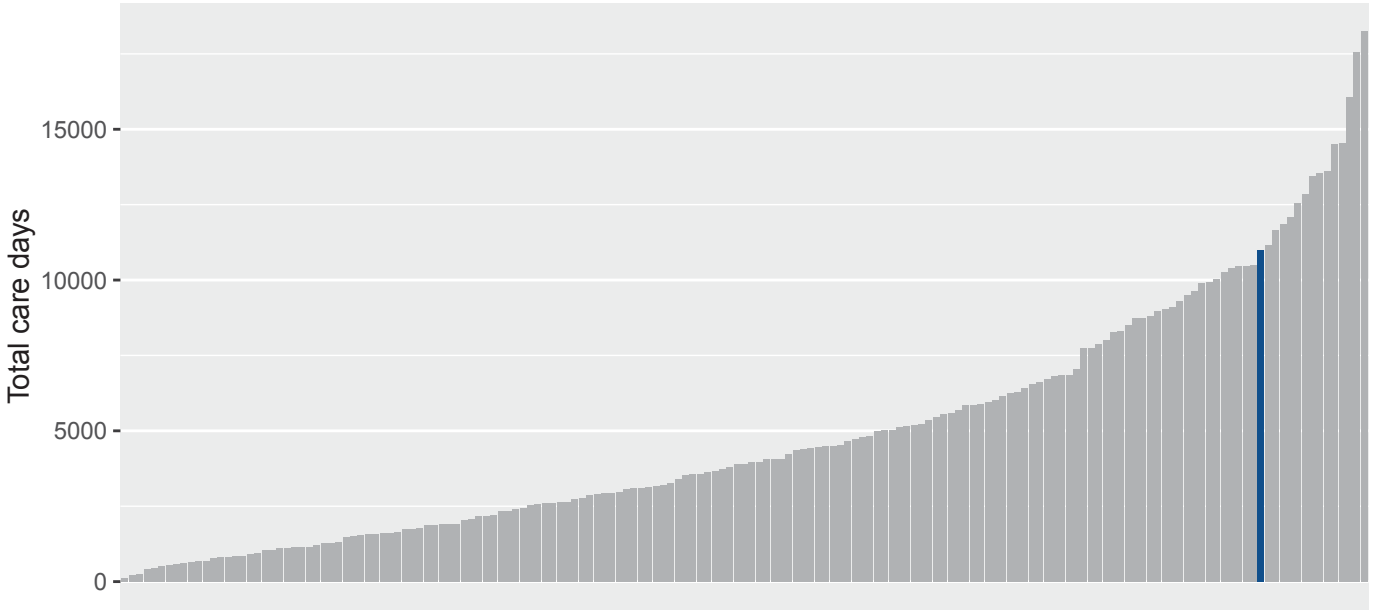


Fig. 8.3 Number of IC days for NNU X (blue) and other NNUs (grey)

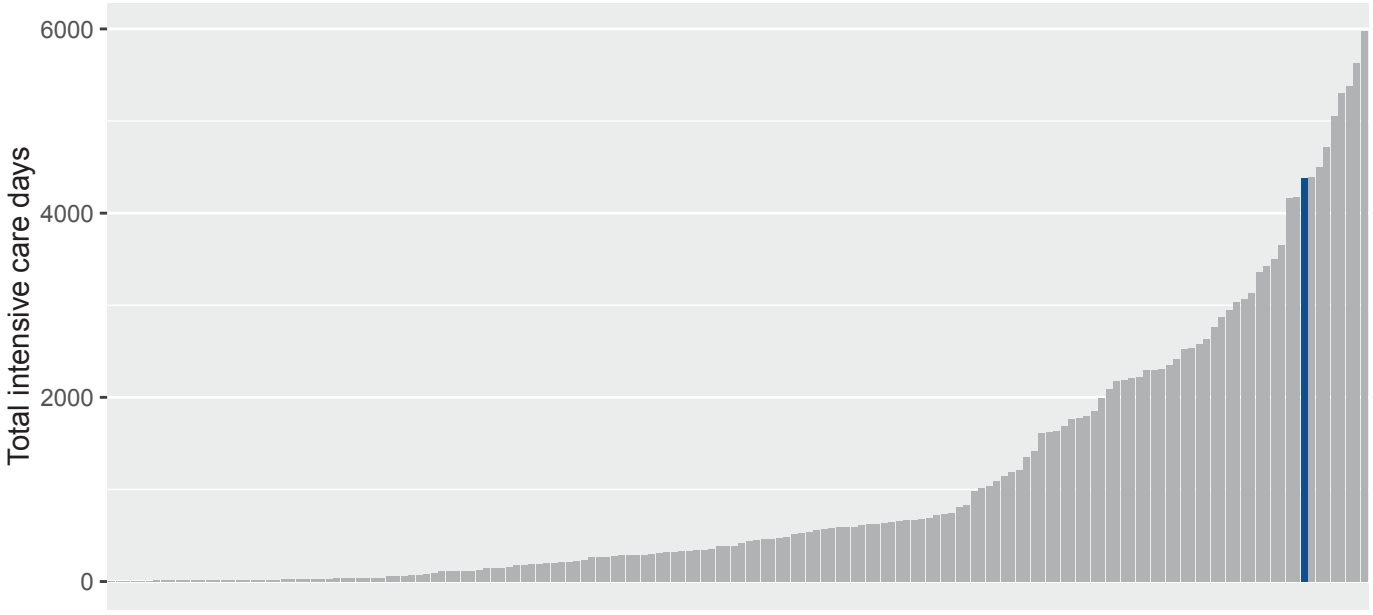


Fig. 8.4 Number of infants admitted to NNU X (blue) and other NICUs (grey)

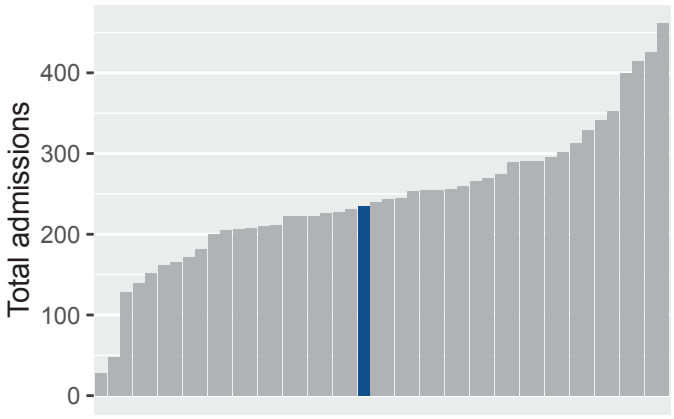


Fig. 8.5 Number of infants admitted to NNU X (blue) and other NNUs in network Y (grey)

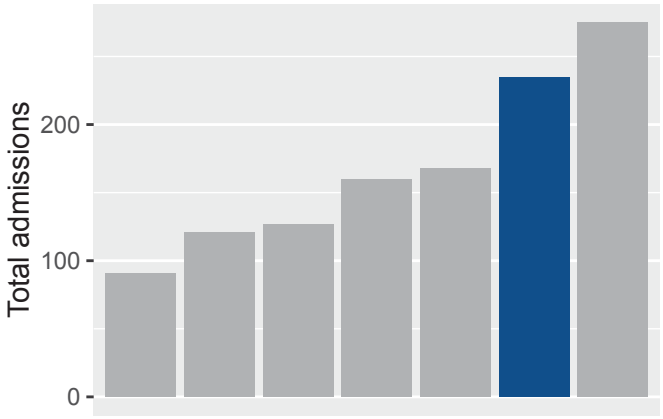


Fig. 8.6 Number of care days for NNU X (blue) and other NICUs (grey)

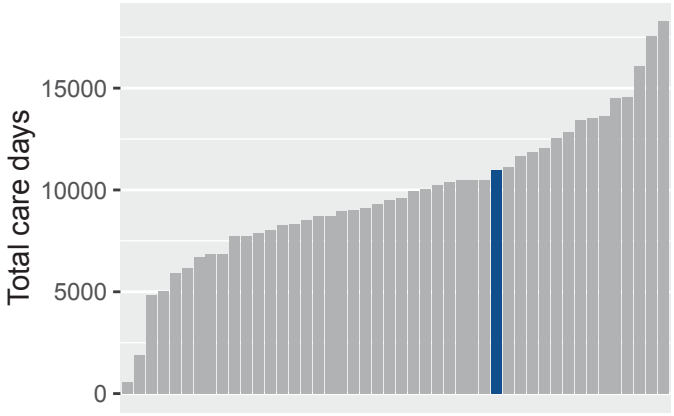


Fig. 8.7 Number of care days for NNU X (blue) and other NNUs in network Y (grey)

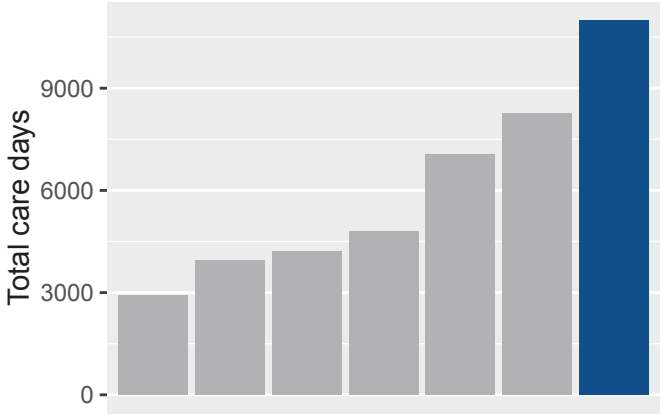


Fig. 8.8 Number of IC days for NNU X (blue) and other NICUs (grey)

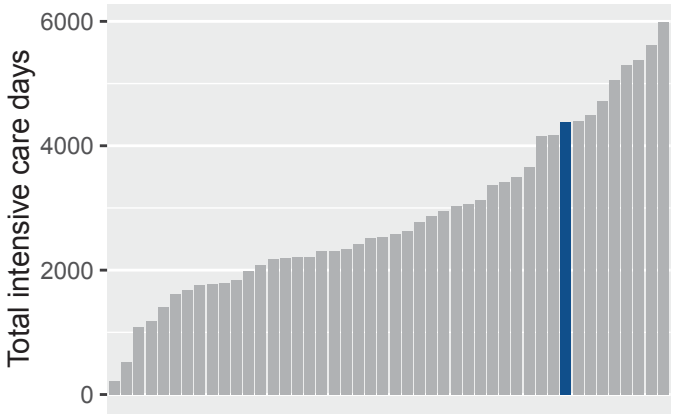


Fig. 8.9 Number of IC days for NNU X (blue) and other NNUs in network Y (grey)

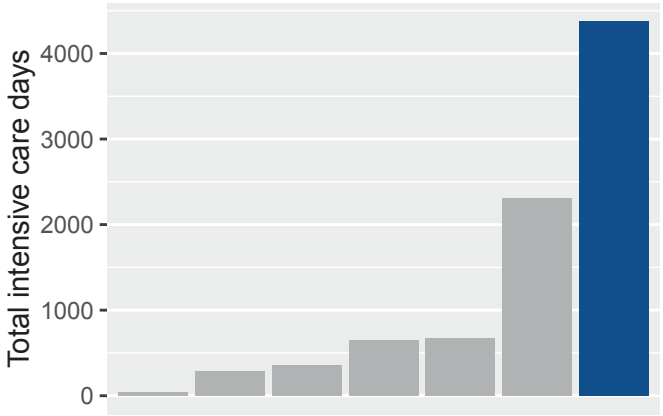


Fig. 8.10 Source of admission for infants admitted to NNU X

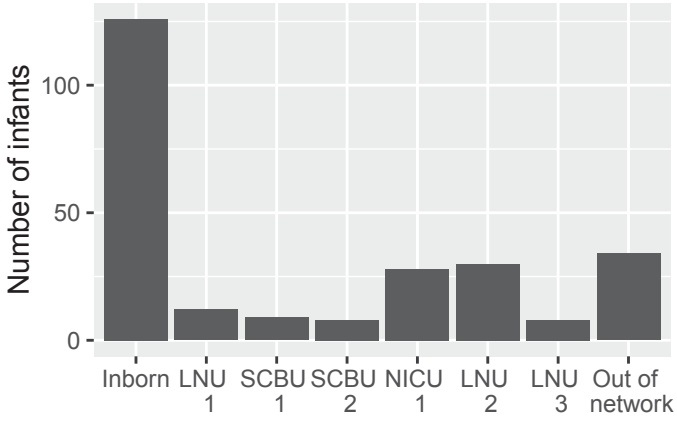


Fig. 8.11 Outcome or transfer destination for infants discharged from NNU X

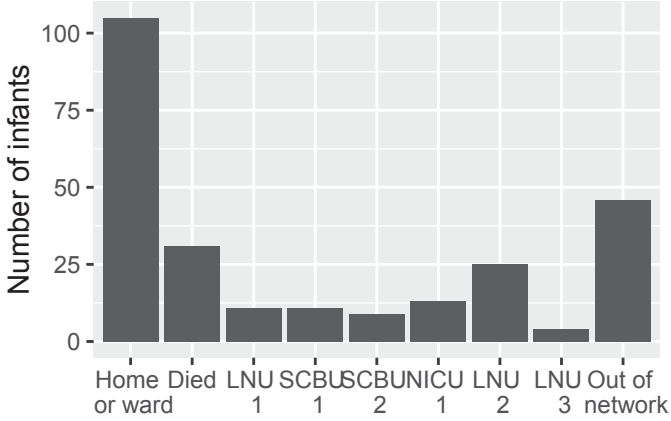


Fig. 8.12 Gestational age distribution for NNU X (blue) and all NNUs (grey)

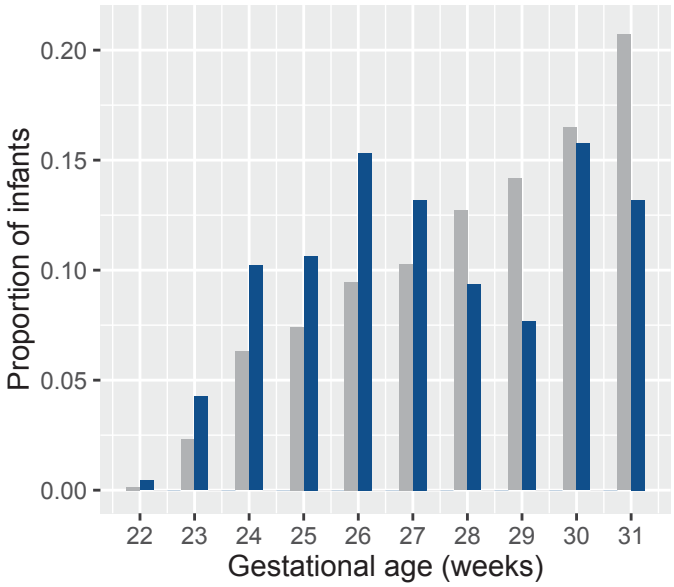


Fig. 8.13 Gestational age distribution for NNU X (blue) and all NICUs (grey)

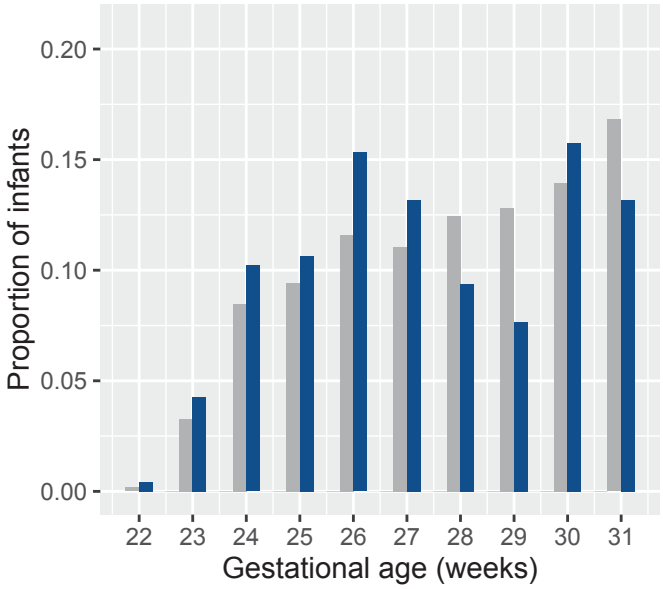


Fig. 8.14 Birth weight distribution for NNU X (blue) and all NNUs (grey

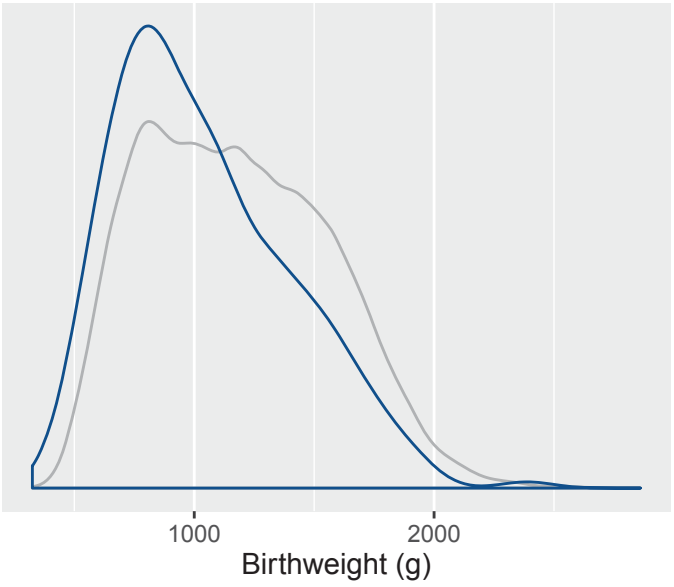


Fig. 8.15 Birth weight distribution for NNU X (blue) and all NICUs (grey)

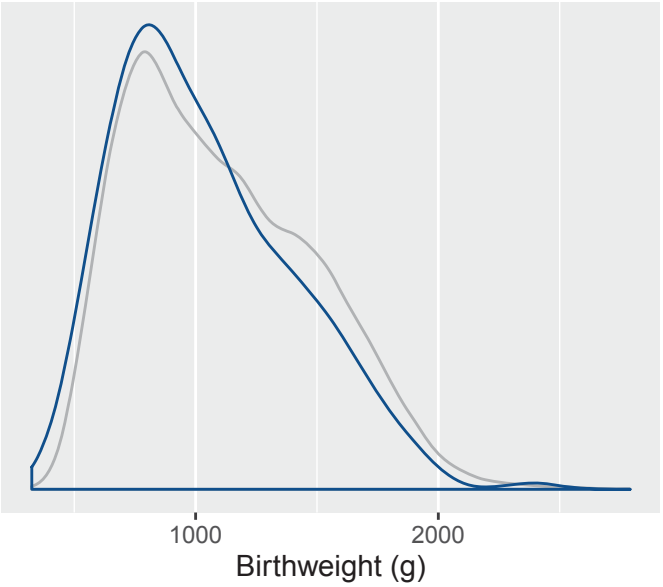


Fig. 8.16 Distribution of predicted risk of death for NNU X (blue) and all NNUs (grey)

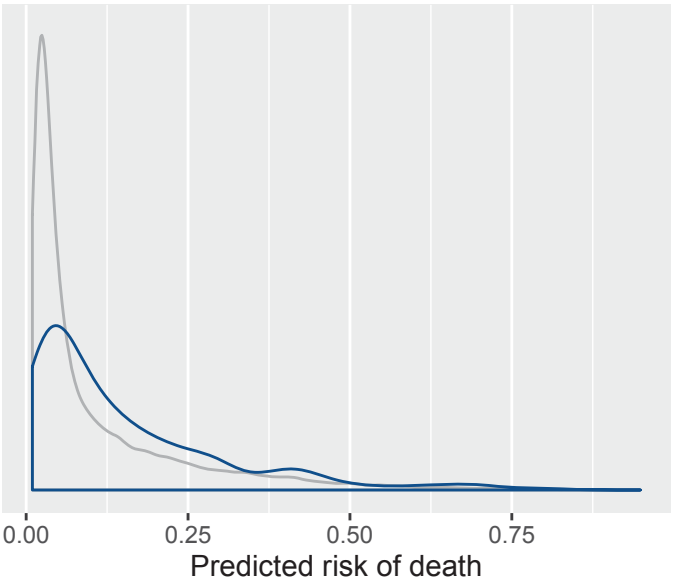


Fig. 8.17 Distribution of predicted risk of death for NNU X (blue) and all NICUs (grey)

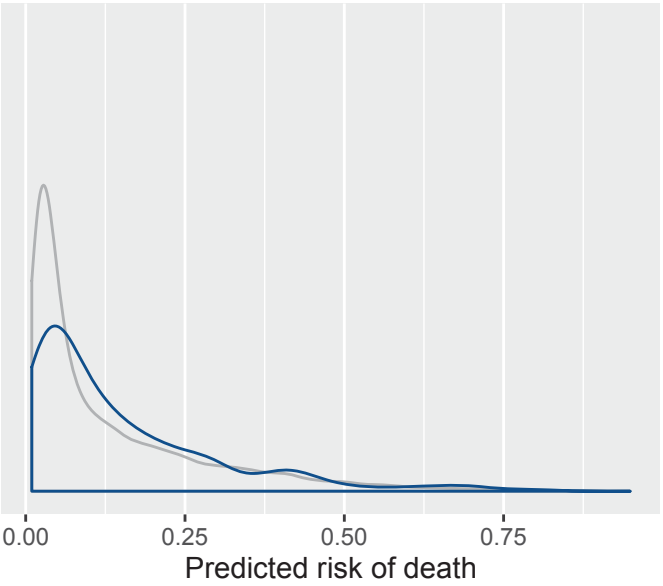


Fig. 8.18 IMD distribution
for NNU X (blue) and all NNUs (grey)

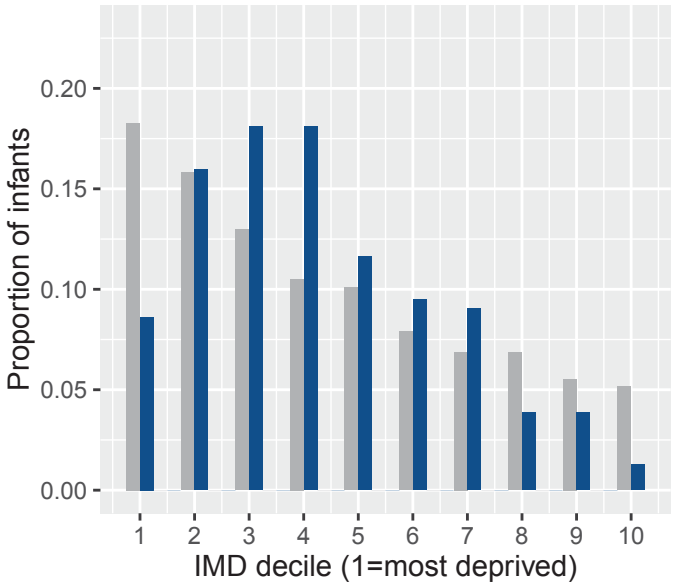


Fig. 8.19 IMD distribution
for NNU X (blue) and all NICUs (grey)

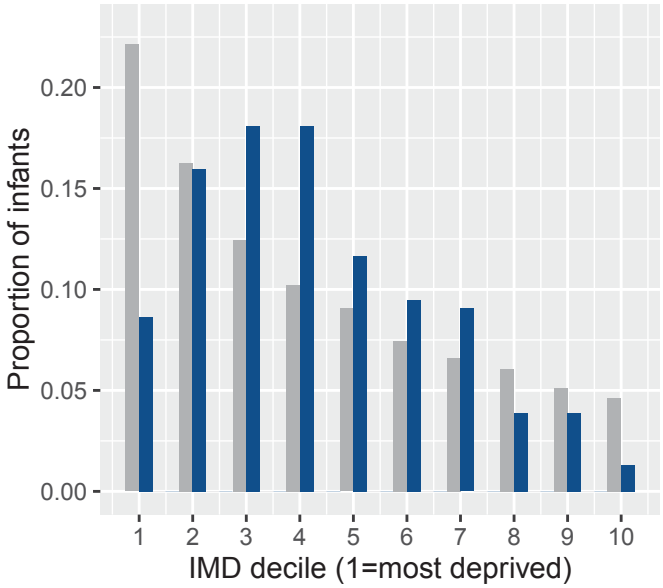


Fig. 8.20 Maternal ethnicity distribution
for NNU X (blue) and all NNUs (grey)

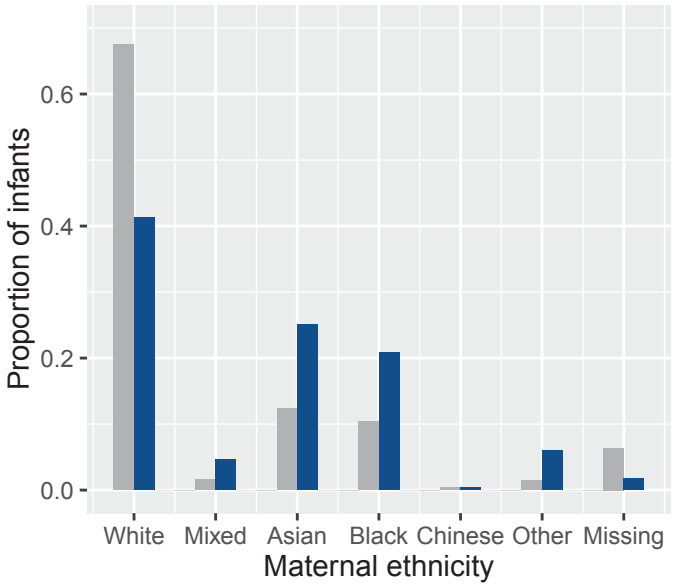


Fig. 8.21 Maternal ethnicity distribution
for NNU X (blue) and all NICUs (grey)

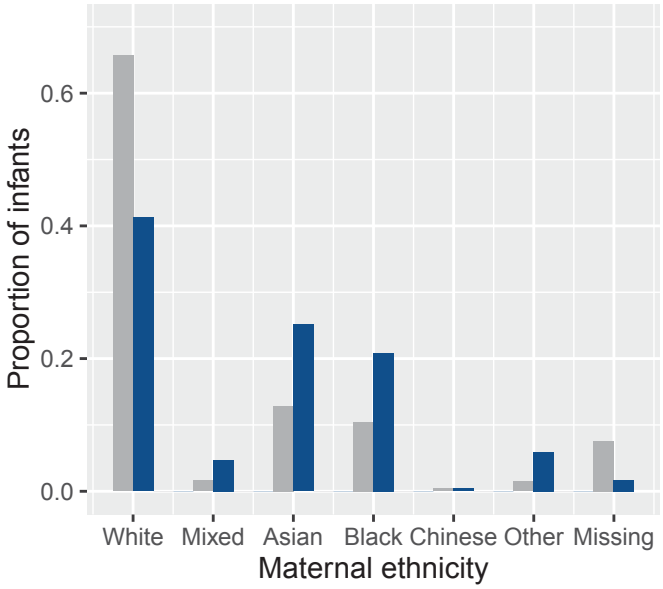


Fig. 8.22 Maternal age distribution
for NNU X (blue) and all NNUs (grey)

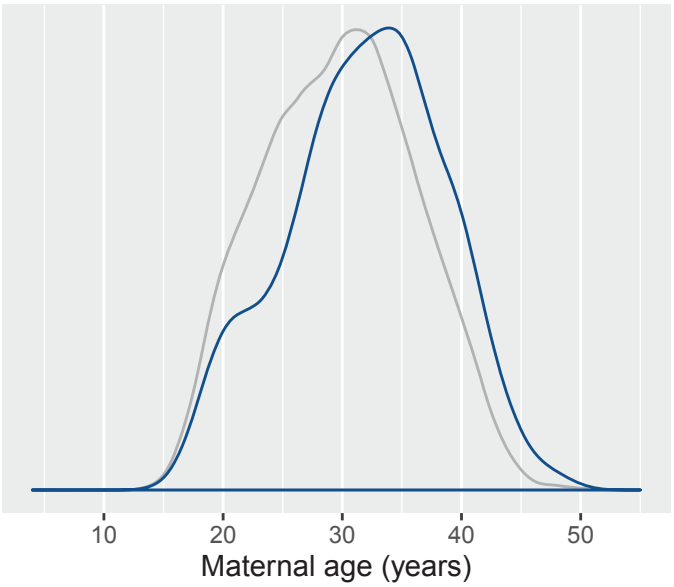


Fig. 8.23 Maternal age distribution
for NNU X (blue) and all NICUs (grey)

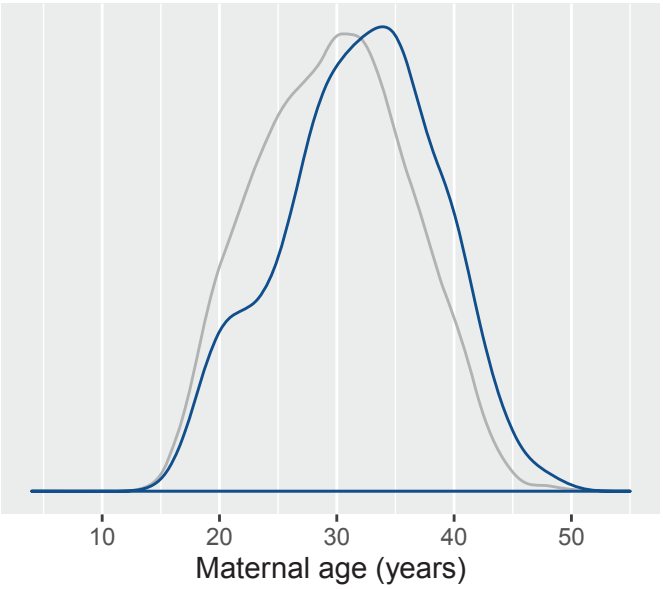


Fig. 8.24 Crude mortality for non−transferred infants at NNU X (blue) and other NNUs (grey)

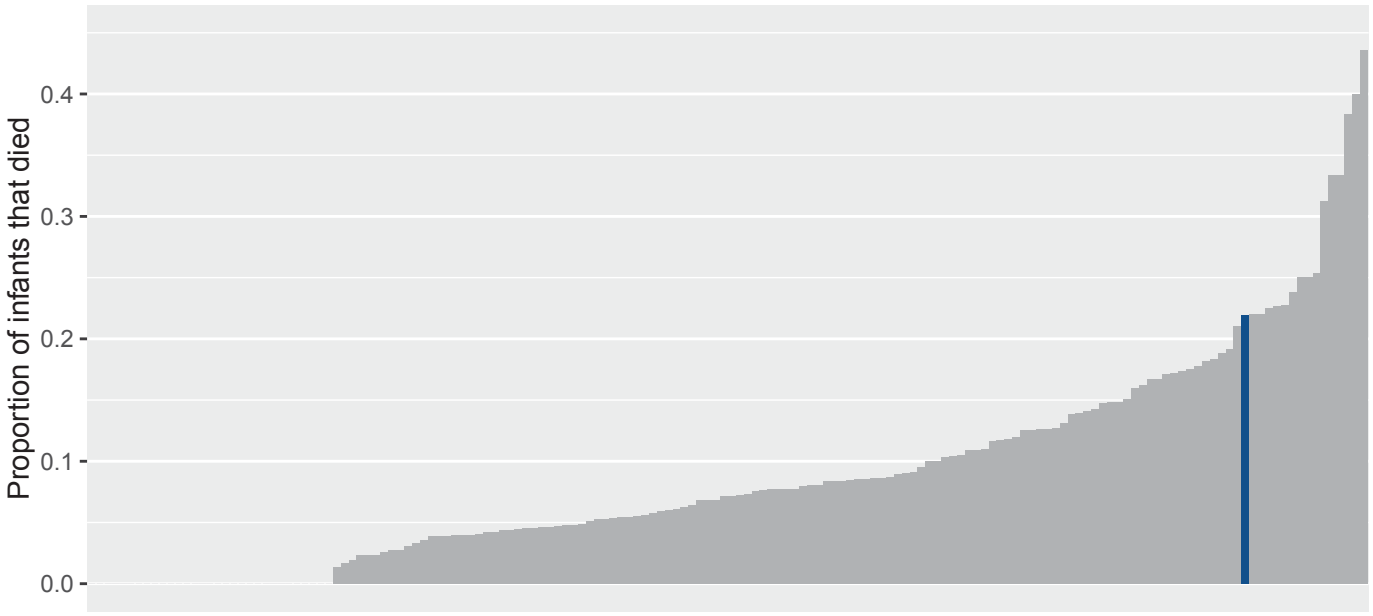


Fig. 8.25 Crude mortality for all infants at NNU X (blue) and other NNUs (grey)

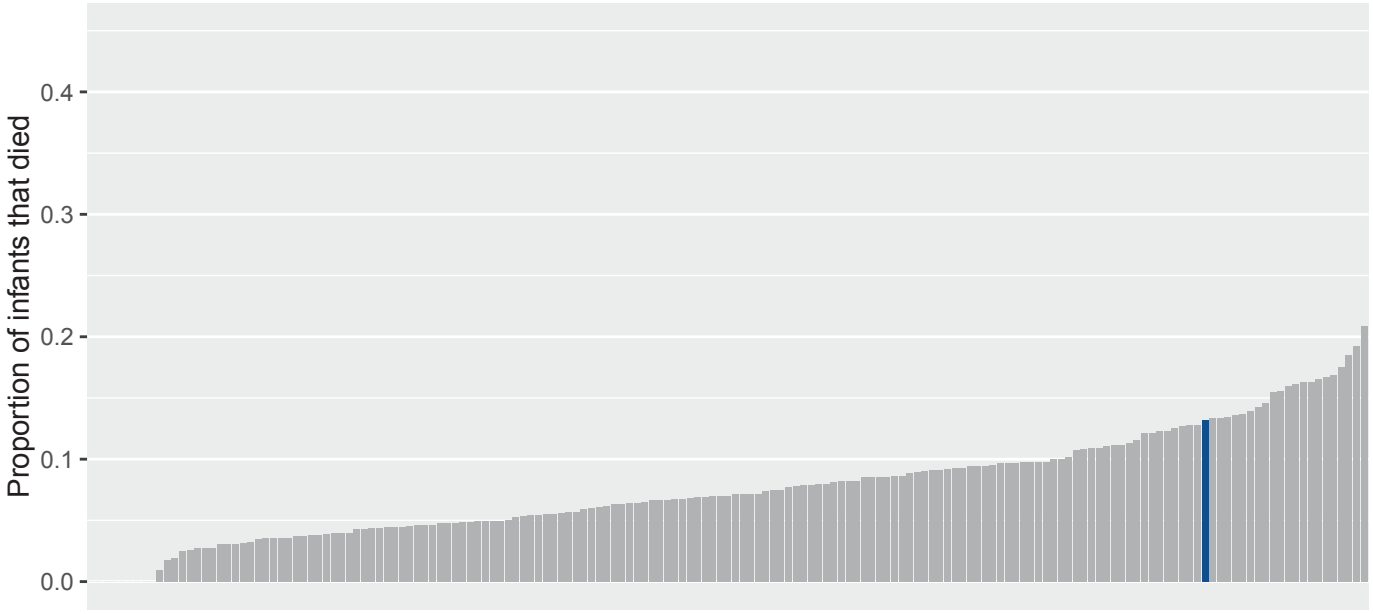


Fig. 8.26 Crude mortality for non−transferred infants at NNU X (blue) and other NICUs (grey)

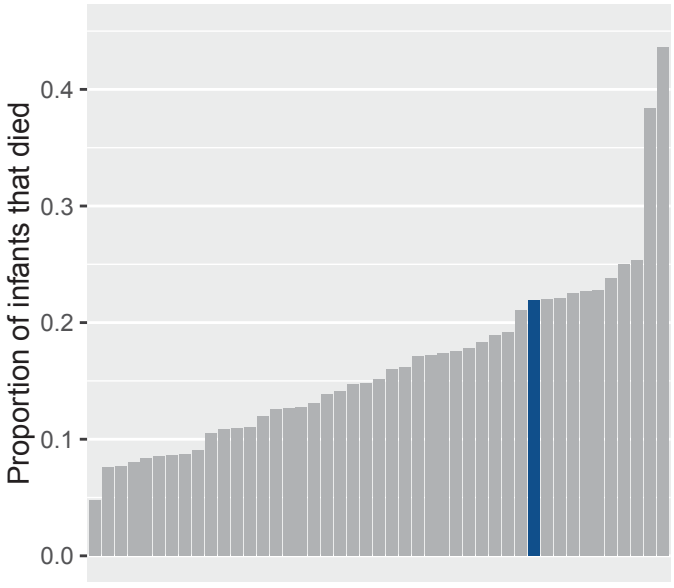


Fig. 8.27 Crude mortality for all infants at NNU X (blue) and other NICUs (grey)

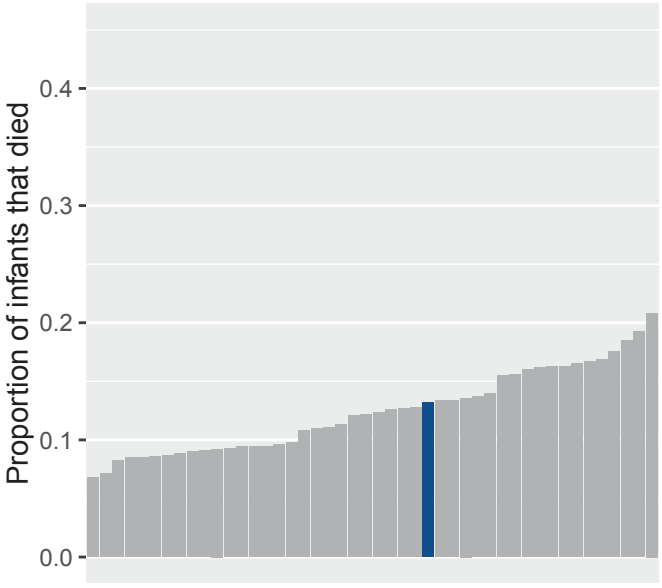


Fig. 8.28 Adjusted OR for non−transferred infants, NNU X (blue) and other NNUs (grey)

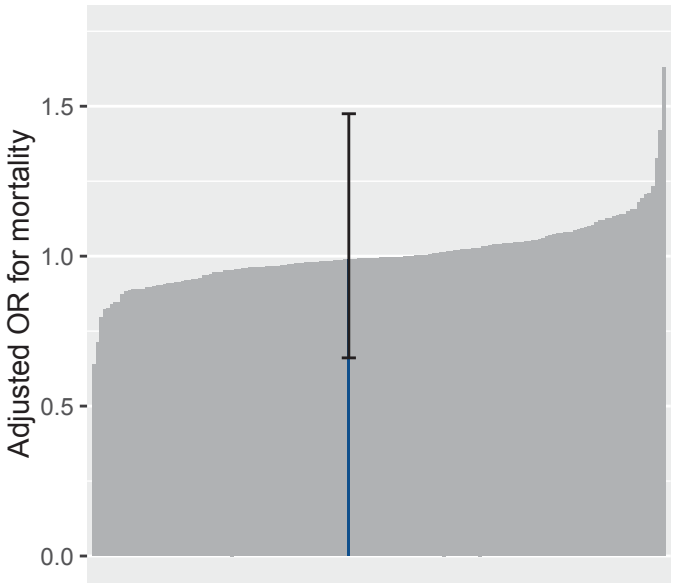


Fig. 8.29 Adjusted OR for non−transferred infants, NNU X (blue) and other NICUs (grey)

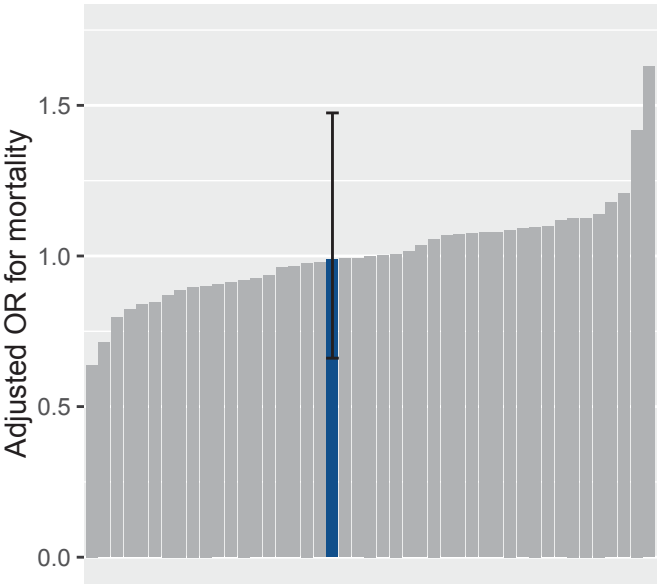


Fig. 8.30 Adjusted OR from equal weight MM model, NNU X (blue) and other NNUs (grey)

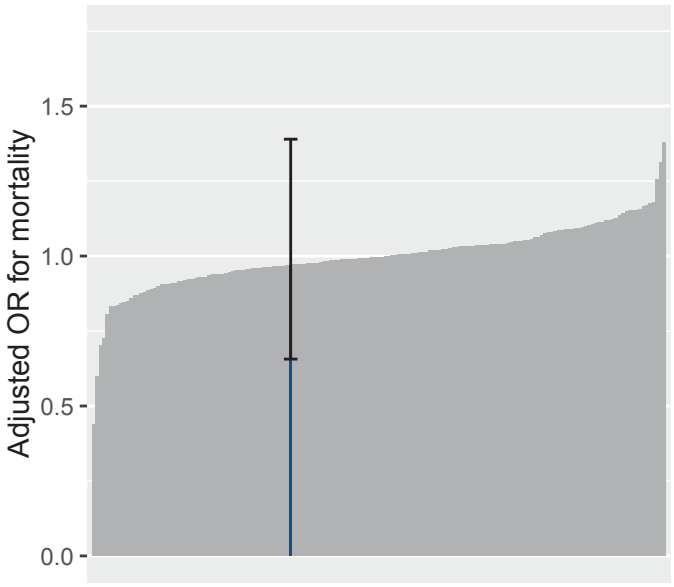


Fig. 8.31 Adjusted OR from equal weight MM model, NNU X (blue) and other NICUs (grey)

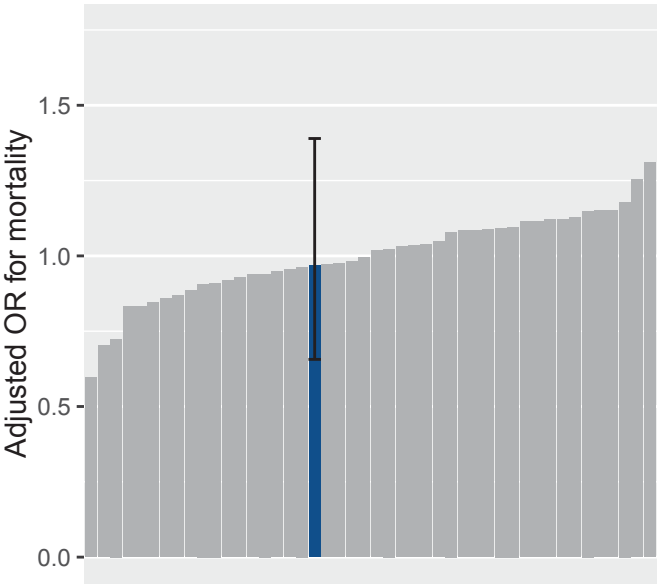


Fig. 8.32 Adjusted OR from weighted LOS MM model, NNU X (blue) and other NNUs (grey)

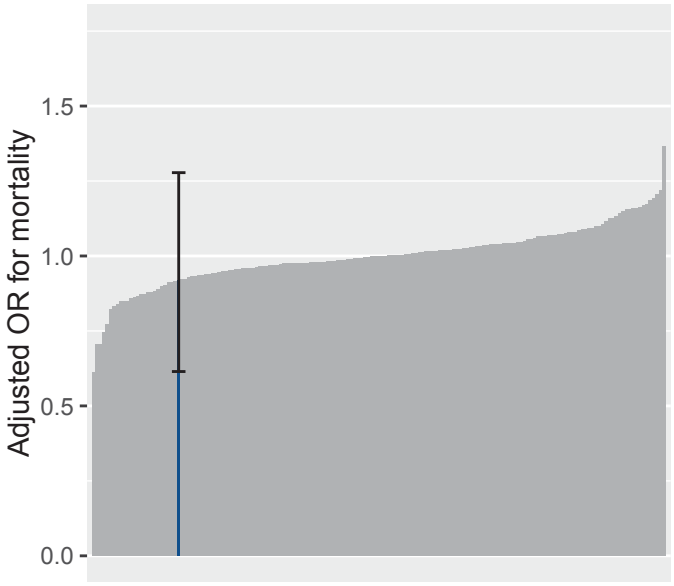
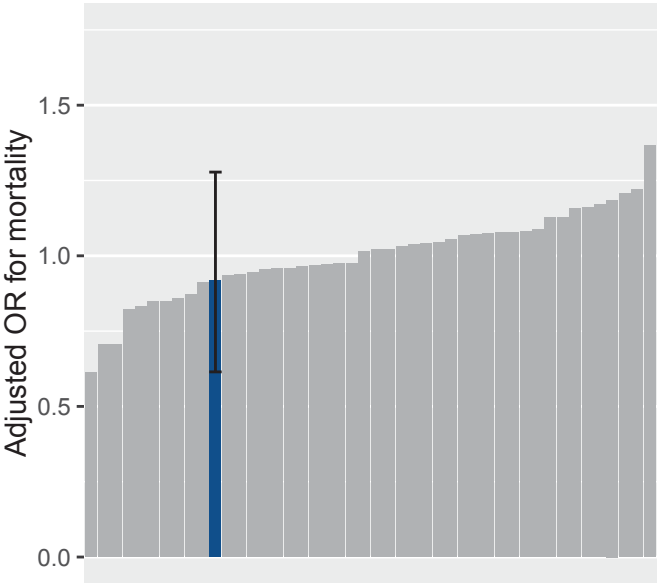


Fig. 8.33 Adjusted OR from weighted LOS MM model, NNU X (blue) and other NICUs (grey)



Chapter 9

Conclusion

In the final chapter of the thesis I begin by revisiting the research question (section 9.1) and reviewing the chapters (section 9.2). In section 9.3 I summarise the contributions of the thesis to the relevant research areas. These are discussed in more detail in section 9.4 including the limitations and relation to existing literature. Further work following from the thesis is discussed in section 9.5 before concluding in section 9.6.

9.1 Research question

Hierarchical models are often used to conduct comparisons of clinical outcomes across healthcare providers (*provider profiling*). Such models acknowledge the dependencies in the data while explicitly modelling the variation that occurs at provider level. In the UK, neonatal care is one speciality where the structure of services is particularly complex. Neonatal units (NNUs) are categorised based on the type of care they offer. In order to provide comprehensive care, the service is delivered via managed clinical networks, with infants transferred between NNUs as required. All NNUs caring for an infant need access to his or her medical records, so for several years clinical data has been captured on a web-based database. This is collated and cleaned to form the National Neonatal Research Database (NNRD), but its use for monitoring mortality rates for individual NNUs has been limited by the complex structure of the data. The aim of this thesis was to establish suitable statistical models to inform NNUs how the expected mortality for infants in their care compares with that of similar infants attending similar NNUs. In particular I investigated assignment of the outcomes of transferred patients, modelling differences across NNU levels, and handling of data dependencies at the neonatal network level.

9.2 Thesis summary

In chapter 1 I described preterm birth, its risk factors and its consequences, demonstrating the increasing importance of this vulnerable population. I motivated the research questions of the thesis by describing the structure of neonatal services in the UK, the factors underlying the development of the network structure, and the nature of postnatal transfers. I gave a background to provider profiling studies, including some of the challenges, before describing the research question and setting out the plan for the thesis.

Chapter 2 described the data source for the thesis (the NNRD), justified the selection of the cohort and presented descriptive analysis of the data. I focussed the research question by defining the cohort and the outcome for analysis. I chose to analyse death before discharge for singleton infants born 2011-2013 between 22^{+0} to 31^{+6} weeks gestational age who were admitted to neonatal care in England. Decisions were based on obtaining a complete and well-defined population for whom a suitable risk-adjustment model for a clinically relevant outcome could be developed. Descriptive analysis showed that the infant characteristics and transfer patterns are broadly consistent with the existing knowledge about preterm birth and neonatal care described in chapter 1.

In chapter 3 I developed a risk adjustment model including gestational age, birth weight, sex and antenatal steroids as factors. Variable selection was based on a review of the literature but restricted to those suitable for provider profiling purposes. The model showed good predictive performance in comparison with existing models. I discussed adjusting for provider-level variables and for postnatal transfer, selecting only NNU level as a factor for inclusion in subsequent analysis.

Existing statistical methods were reviewed in chapter 4. I described the suitability of hierarchical models, in particular Bayesian techniques, for provider profiling generally and for the thesis. These methods were shown to be suitable for modelling variation at the neonatal network level and differences in levels of NNU, but not neonatal transfers. A systematic review of the literature for approaches to handling transferred patients showed that all studies either excluded transfers, allocated to a single provider or did not use hierarchical models. I then proposed multiple membership (MM) models as a potential solution and described some existing applications and methodological findings relevant to the thesis. In particular I described considerations with regard to choosing MM weights and found that little attention has been paid to estimating the random effects themselves, which is the focus of this thesis.

Chapter 5 sets out the methods for the thesis. As a preliminary I gave an overview of the Bayesian techniques that would be employed in the chapter. I then described a

series of models to investigate modelling variation by NNU level and neonatal network, applied to non-transferred infants only. This was followed by a simulation study to assess how sensitive parameters of interest were to a variety of MM data scenarios and analysis models, proposing some alternatives to the standard weight which might be more suitable for neonatal data. Finally I described the application of MM models with a series of different weights to the neonatal data, building on the models developed for the non-transferred infants.

Results for non-transferred infants were presented in chapter 6. There were differences in mortality between NNU levels with both special care baby units (SCBUs) and neonatal intensive care units (NICUs) having higher adjusted odds of mortality than local neonatal units (LNUs). There was some indication of higher variation for SCBUs and NICUs compared with LNUs, but variance estimates were uncertain. As this led to less borrowing of information for smaller providers and greater uncertainty in the resulting estimates, a common between-NNU variance was modelled instead. There was little evidence of variation between neonatal networks but modelling this allows us to obtain estimates of network effects, which is important given that they are also providers.

In chapter 7 I presented results from the MM analyses. The simulation demonstrated that assigning weight too unequally across providers (with assignment to a single provider being the most extreme case) underestimated the between-provider variance. Assignment to a single provider produced provider effects which had lower uncertainty but poorer coverage, with less difference seen across the MM models. These results were echoed in the analysis of neonatal data, but differences were larger. The different aspects of the analysis produced a consistent pattern across the weighting schemes, with the standard MM weighting of allocating in proportion to length of stay (LOS) producing the largest differences between NNU levels and the largest NNU effects. When and where an infant is transferred is dependent on his or her health status. MM weights that place greater emphasis later during stay, when there is greater separation in the location of care for infants that survive to discharge and infants that die, produce larger differences. These differences persist after adjustment for NNU level, suggesting there are differences in care pathways which are not completely characterised by NNU level.

Chapter 8 showed how the results arising from chapters 6 and 7 might be presented graphically alongside other descriptive data from chapter 2 in a comparable way as a case study for a single NNU.

9.3 Contributions of the thesis

In this section I briefly summarise the novel findings from the thesis which will be discussed further in section 9.4.

Neonatal care In this thesis I have developed methods to evaluate mortality of preterm infants at individual NNUs across England, which has not been done before at a national level. I have showed that there is a low level of residual variation in mortality across NNUs and networks after controlling for key infant characteristics and level of NNU. The risk adjustment model developed as part of this process showed improved model performance compared with three existing models currently in use. Descriptive analysis of transfer patterns and differences in care provided across levels of NNU provides insight into the functioning of the network system, and its impact on the burden of care across levels of NNU. A case study for a single NNU showed how the descriptive data could be presented alongside the mortality analysis to aid interpretation.

Provider profiling I have demonstrated a statistically principled way of accounting for transfer between hospitals in provider profiling studies which has not been done before either in neonatal care or in other specialities. Hierarchical Bayesian approaches are an established method for provider profiling; this thesis adds to this body of work as a practical application of evaluating providers at two levels (NNUs and networks), with providers of different types (levels of NNU) in the context of centralised care. It also shows the utility of using electronic patient records for provider profiling purposes.

Methodology This thesis represents a new application of MM models for modelling patient transfers, including transfers to a different network. I used Beta distributions with different parameters as a family of weights for MM models, which to my knowledge has not been done before. The simulation study demonstrated that the bias and coverage for the random effects was poorer when the MM structure was ignored. Results suggested a pattern regarding the underestimation of the cluster variance and the MM weights used, providing support to the use of equal weights. Application of MM models to the neonatal data revealed that results are more sensitive to choice of MM weights when the pattern of membership is associated with the outcome.

9.4 Discussion

9.4.1 Neonatal care

NNU effects

Mortality of preterm infants at individual NNUs in England has not been evaluated at a national level before. In this thesis I demonstrated that hierarchical multiple membership models yield NNU-specific estimates of risk-adjusted mortality in the form of odds ratios, allocating infants to all NNUs which have cared for the infant, and incorporating neonatal network effects and differences across types of NNU. Mortality at the neonatal network level has been evaluated using Standardised Mortality Ratios (Neonatal Data Analysis Unit, 2015), and other outcomes at the NNU level for the National Neonatal Audit Programme (Royal College of Paediatrics and Child Health, 2015), where transfers were either excluded or assigned to one NNU. Hierarchical models have been used to profile neonatal providers elsewhere (Bonet et al., 2011; Lee et al., 2013; MacNab et al., 2005), though again transfers were assigned to a single provider. The differences in approach between those studies and this thesis were described in more detail in section 4.4.2. While this thesis has established the means for calculating and reporting mortality data for NNUs, all admissions should be included before it can be implemented. In particular, the methods shown here should be extended to include infants with missing data and those from a multiple pregnancy. This will be discussed as part of further work in section 9.5.

The inferences drawn about the NNU effects depend on the MM weightings used (see section 9.4.3). That said, NNU effects themselves are fairly robust to different specifications for MM weights, given the overall uncertainty. NNU effects should also be interpreted in light of other data about the NNU which, for reasons discussed in section 3.2, were not included as risk adjustment variables. To illustrate this, in chapter 8 I showed how the results could be presented in the context of descriptive data in a case study for a single NNU.

Risk adjustment

While the risk adjustment model developed here did not use any novel methodology, there were no other models in the review by Medlock et al. (2011) (the basis for variable selection in chapter 3) which used restricted cubic splines to model the effect of gestational age. While discrimination was similar, the model for this thesis gave improved prediction over three existing models ((Manktelow et al., 2013; Parry et al., 2003; Tyson

et al., 2008)), all of which over-predicted mortality. This highlights the importance of re-calibrating models using recent data to keep current with improvements in neonatal survival. This is made feasible by using the NNRD, as it is an ongoing and regularly updated data source. Estimates of gestation-specific mortality showed an improvement from 2006 estimates reported by the EPICure 2 study (Costeloe et al., 2012) for neonatal admissions born 22^{+0} and 25^{+6} weeks gestation. Data were consistent with recent estimates for livebirths from the UK Office of National Statistics, supporting the validity of the findings.

Transfer patterns

Although not a primary aim of the thesis, the descriptive analyses of patterns of care and transfer described in chapter 2 are interesting findings which build on other national (Gale et al., 2012b; Redshaw and Hamilton, 2006) and regional (Cusack et al., 2007) studies, while the detail regarding type and timing of transfer has not been reported at a national level before. A limitation of the analysis is that causal inferences cannot be drawn from the differences between levels of NNU described in chapters 6 and 7. There are several studies (Binder et al., 2011; Marlow et al., 2014; Neto, 2006; Phibbs et al., 2007; Watson et al., 2014) providing evidence of better clinical outcomes for preterm infants born at hospitals with neonatal intensive care facilities. The associations found here suggest higher mortality at NICUs compared with LNUs even after risk-adjustment. In sections 6.6 and 7.5 I argued that this is likely to be due to a selection effect as they treat higher risk infants and conduct more intensive care. The selection effects are amplified for transferred infants because the condition of an infant determines where they are transferred to (this is also observed within level of NNU, as will be discussed in 9.4.3). These differences in case-mix were not captured by the risk adjustment. For non-transferred infants, exploration of factors known before birth which determine whether an infant requires planned delivery at (or in-utero transfer to) a NICU could yield other variables to improve risk adjustment. For transferred infants, a different method which can model risk factors that change over time is required; possibilities for this will be discussed in section 9.4.3.

9.4.2 Provider profiling

This thesis showed that provider profiling analyses can accommodate patient transfers in a hierarchical modelling framework without assigning to a single provider or undertaking multiple analyses to compare assignment to different providers. The approach taken to accommodate transfers is similar to that of Hassani et al. (2015) and Kristoffersen et al. (2012) in that outcomes for transferred patients are shared across all providers.

Those studies used fixed effects regression, with Hassani et al. (2015) shrinking the provider effects in a second stage. Using MM models means the provider effects can be estimated in a single analysis. Several studies in the systematic review (section 4.2) advocated allocation to the referring provider. This thesis showed that following this recommendation is not suitable if we want the provider effects to reflect the whole of stay as it underestimates the between-provider variance and results in provider effects which are over-shrunk, with credible intervals that have poor coverage of the true effects. In addition, this recommendation as a general rule is unsuitable for neonatal care (and possibly for other networked specialities), as there are multiple referring providers for infants transferred more than once. If taken to mean the first provider, it would result in many transferred infants assigned to an NNU providing less than 24 hours of care.

The hierarchical Bayesian models applied here are an established method for provider profiling which is suitable for settings with many small providers (DeLong et al., 1997; Goldstein and Spiegelhalter, 1996; Normand et al., 1997). Some analyses have adjusted for type or provider in some way to enable peer group comparison (Austin et al., 2004; MacNab et al., 2005), with others allowing for complex variation by type of provider (Goldstein, 2002). Extensions of basic models have been implemented to accommodate a third level of dependence (Brookhart et al., 2006; Chen et al., 2013) but not to my knowledge in the context of networked care, where the third level presents a co-ordinated group of providers at the lower level. The thesis also serves as an example of using electronic patient records for provider profiling, which is advantageous as it minimises additional costs, including staff time for data entry. Some practical aspects of using routine data require further attention: though some validation processes are in place (section 2.1.3) no further validation was conducted specifically for the thesis, nor was the potential effect of data errors on the results investigated. Infants with missing outcome or covariate data were excluded; further work on handling missing data is described in section 9.5.

Provider profiling studies are often focussed on detecting unusual performance and comparing providers (Goldstein and Spiegelhalter, 1996; Ohlssen et al., 2007b). Shrunken provider effects arising from hierarchical models are often used for these purposes, though the limitations regarding power and inconsistent shrinkage are known (Aylin et al., 2003; Shen and Louis, 1998). As discussed in section 4.4, the results of this thesis are not suitable for these purposes. Birth below 32 weeks gestation is rare, and subsequent mortality more so, therefore power to detect unusual performance is limited, even at the larger NNUs. There was considerable variation in the size of NNU (both in general and in terms of the number of high risk infants), so the level of shrinkage varied across NNUs, limiting the utility and fairness of comparison and outlier detection. I have instead focussed on estimation of NNU-specific mortality. However, as seen in the caterpillar plots in sections 6.5 and 7.4.5, uncertainty around the provider-specific estimates is high, par-

ticularly for smaller NNUs. This is an unavoidable limitation given the numbers involved for this particular population and outcome. This could be mitigated by examining data over a longer period of time (perhaps on a rolling basis to enable regular assessment to be conducted) but a drawback is the loss of currency. Models and weights where more information was borrowed across NNUs yielded more precise results. This could be increased by including more provider-level variables (e.g. co-location of surgical facilities, availability of a specialist transfer service), as this reduces the residual between-NNU variability by refining the peer group to which NNUs are compared.

9.4.3 Methodological

This thesis presents a new application of MM models, motivated by their use in modelling pupil mobility in education (Goldstein et al., 2007; Leckie, 2009), which is a similar problem. Neonatal transfers are an interesting application for multiple membership models with features that may be useful to others applying MM models.

Firstly, data on the intensity of care are available to inform weights, allowing for weights to vary by individual. Secondly, episodes of neonatal care are not of discrete duration and total length of stay varies considerably across infants (as is likely to be the case for patient transfers in general). This led to the use of the Beta distribution to specify MM weightings. While this was not finally selected as the most appropriate method for this application, it may prove useful in other applications of MM models. It allows the analyst to describe a family of weights with different shapes, so could be used to check sensitivity to assumptions.

Finally, in neonatal care the location of care at a point in time is indicative of the infant's outcome, so there is the potential for a stronger relationship between weighting patterns and clustering effects which was made evident in chapter 7. This has implications for the choice of MM weight, unlike many of the existing MM studies which found results were robust to choice of weight (Fielding and Yang, 2005; Goldstein et al., 2006; Leckie, 2009). I argued in sections 7.5.3 and 7.5.4 that for provider profiling purposes the choice of weighting should be driven by the inferences we wish to make, so equal weight is a reasonable choice as it has an intuitive interpretation of equal accountability. For neonatal care, and other settings exhibiting similar patterns, equal weight is appealing for its neutrality and simplicity, but sensitivity analysis using different weightings is important, giving us a range of results with different inferences. In neonatal care, the weighted LOS is a useful comparison as it mitigates some of the problems of the standard LOS weighting (see section 7.5.3).

Logistic regression treats death as a single binary outcome, and therefore we only

have a single set of risk factors measured at birth. In section 4.4 I considered discrete time logistic regression models as a possible alternative which models the outcome at the end of each episode as a set of sequential, dependent outcomes. These would allow the condition of the baby at the start of each episode to be considered separately (if a suitable set of risk factors can be found) which, in light of the findings from chapter 7, might help resolve the issue of divergent pathways for sicker patients. However I am unaware of any hierarchical implementation of these models where individuals have different numbers of episodes of varying lengths, as is the case here. Another possible alternative is to treat the outcome as a survival time. While this is not a standard approach for perinatal mortality given the short timescale, it would allow the risk of dying to change over time. Random effects models for survival data are known as *frailty* models. Transferred infants whose cluster membership changes could possibly be modelled by a time-dependent frailty term, though these are primarily used for recurrent events (Manda and Meyer, 2005; Yau and McGilchrist, 1998). I have not explored these methods further as part of the thesis, but their complexity in contrast to MM models may limit their utility in routine reporting of clinical outcomes.

It is already known that cluster variance is underestimated when the MM structure is ignored (Chung and Beretvas, 2012; Fielding and Goldstein, 2006; Goldstein et al., 2007; Leckie, 2009). The simulation study here indicated that the underestimation is also present when the weights applied in the MM analysis are more unequal than the process generating the data. This is in line with findings from Wolff-Smith and Beretvas (2014) but results here also suggest greater underestimation for more unequal weights. I also showed that the random effects were more biased towards the null, and uncertainty was underestimated resulting in poorer coverage. Previous studies comparing MM weights and single membership models have not examined these issues. The simulation study here was limited and further exploration with a wider range of parameters and more complex data is required to solidify these findings. In particular, further support for the proposition that assuming equal MM weight results in minimal bias would strengthen the case for using equal weights.

9.5 Further work

9.5.1 Missing data

Infants were excluded if they had missing covariate or outcome data. Having developed methods suitable for complete case data I consider how these could be extended to include missing outcome data. The Bayesian Markov Chain Monte Carlo (MCMC) methods applied here are well suited to missing data problems because our uncertainty

about the missing data is no different to the uncertainty about any other parameter, and possible values for the data are imputed with each MCMC iteration (Gelman et al., 2004). Were the infants with missing outcomes no different to those with complete data, the logistic regression model would already capture all our knowledge about the outcome. We can use this same model to impute a posterior distribution for the missing outcomes (following the addition of a suitable prior for the missing data to the model formulation).

In section 2.2.7 I discussed the reasons for missing outcomes and the resulting statistical implications. Briefly, the most plausible reason for missing outcomes is that infants have been transferred to specialist surgical providers. The fact that the outcome is missing is likely to be associated with the outcome itself: the need for surgery is an indication of illness severity, and the surgery itself will affect the outcome. Therefore a model for the missingness mechanism is required. Similarly to constructing the hierarchical models by considering each level separately (infants, NNUs, networks), so we can have a separate "submodel" for the missingness mechanism, with all parameters estimated simultaneously (Mason et al., 2010). We model the probability of missingess as a function of the outcome and any other information we wish to include, though careful thought is required as to how this could feed back into the provider effects. Such models usually require informative priors for parameters of the missingness model, with a range of priors applied to reflect different assumptions regarding the missingness mechanism.

Even if we can model the data and missingness in suitable ways, a further question is how the outcome for these infants should be attributed to the NNUs caring for the infant. Outcomes for these infants could be attributed as if the (imputed) death or discharge home occurred from the final NNU. This ignores the role the specialist unit plays in the survival of the infant, which is possibly a crucial one. Specialist units are not amongst the providers being profiled as they are not neonatal units. There is currently too little data to include them in any modelling, though a possible future development is that surgical providers will input data into the electronic data capture system.

9.5.2 Multiple births

The cohort for the thesis consisted of singleton infants only who, while a distinct and clinically relevant population in their own right, comprise only part of the admissions to neonatal care (73% of infants in the NNRD born 2011-2013 below 32 weeks gestation). Including multiple births poses two areas of difficulty. Firstly, infants from multiple pregnancies are different from their singleton counterparts and therefore may require a different risk adjustment model (see section 2.4 for further discussion). Secondly, they result in additional correlations in the data because infants in the same multiple birth

set are likely to have similar outcomes beyond the similarity in measured risk factors such as gestational age. The extent of these correlations requires investigation, but as described in 2.2.5 multiple birth sets could not be adequately matched in the NNRD data. The structure of multiple birth data is problematic because the clusters (multiple birth sets) are small, which can lead to estimation problems, and because not all the data are clustered because singleton infants are not part of a set. Generalised estimating equations are often used in the twin literature but they cannot be combined with hierarchical models (Carlin et al., 2005; Marston et al., 2009). While ignoring partial clustering in the term population when the proportion of multiple births is low may be reasonable (Marston et al., 2009) this is not the case for preterm infants. Furthermore, as it is the clustering at NNU level which is of interest and multiple births are usually treated in the same NNU (Gale et al., 2012b), any unexplained variance from multiple birth sets will be attributed to NNU level, resulting in over-estimation of the between-NNU variance.

Partially clustered data are frequently seen in psychotherapy trials (Baldwin et al., 2011). Patients receiving treatment from a psychotherapist are clustered within therapists, (as we would expect there to be some correlation due to the subjective nature of the treatment) while those in a control group or receiving self-administered treatment would not be clustered. One approach is to model an interaction between the clustering variable (therapist treatment arm or, in our case, multiple birth indicator) and the random effect, so that the clustering effects are only modelled for the clustered data. For psychotherapy trials this is intuitive as only those receiving therapy have a variance contribution from the therapist effect. This is not the case when considering infants and multiple births. The similarity between twins will be due to the characteristics of the parents, the pregnancy and the birth. Singletons will also be subject to these influences so it may be more sensible to model them as a cluster of size one. Sauzet et al. (2013), motivated by the high proportion of multiple births in the preterm population, showed that hierarchical models produced unstable parameter estimates for continuous outcomes. Additionally they found that the random effects variance was underestimated when the number of clusters was small, when there were many predictors or where there was less heterogeneity between clusters. The models developed in this thesis can in theory be extended to accommodate multiple births using partial clustering at the infant level, but a simulation approach in the manner of Sauzet et al. (2013) is advisable to check sensitivity.

The proportion of preterm infants from a multiple pregnancy is substantial and possibly growing, as the rise multiple birth rates is thought to be a key drivers of increases in preterm birth. The reasons for early delivery are not the same as those for singletons and their clinical needs may also differ. Including multiple births is therefore a priority for development of this work.

9.5.3 Other applications

Death before discharge was chosen as the outcome of interest for this thesis as it is clearly defined, well-recorded and of the utmost clinical importance. Very preterm birth and subsequent mortality is, fortunately, relatively rare in the UK, meaning that there is little power to determine with certainty whether the mortality at an individual NNU is particularly high or low. In section 1.1 I mentioned other adverse consequences of preterm birth. Of these bronchopulmonary dysplasia (BPD) is well-defined and recorded in the NNRD, occurs with higher frequency, and affects infants born at later gestations (32 to 37 weeks) who form the majority of the preterm population. Daily data on the type of respiratory support received by each infant is well-recorded in the NNRD, so could be used to inform the MM weights. A UK-wide survey regarding neonatal resuscitation practices by Mann et al. (2012) revealed variations in clinical management, particularly by NNU level. If there are specific policies which are of interest, these could be considered for inclusion as provider-level variables. Examination of BPD using the methods developed in this thesis may help determine if such variations have clinical consequences.

The NNRD is being employed more extensively beyond service evaluations, audits and commissioned studies. Of particular note is ongoing research into using the NNRD for clinical trials (Gale et al., 2016a). While the focus is not on estimating random effects as in this thesis, appropriate handling of dependence is required for unbiased and accurate estimation of treatment effects. The methods used in this thesis may have applications in the analysis of multi-centre trials using the NNRD. Utilising routinely collected health data for pragmatic trials is growing in popularity due to the gains in efficiency it can bring to the extent that trials which would otherwise be too costly and too much of a burden to run become feasible. These types of data are often collected in settings with complex patient pathways (e.g. GP referrals, multiple hospital episodes) so there could be uses outside neonatal medicine as well.

There is growing evidence of the benefits of focussing resources at a few key specialist centres rather than attempting to offer all services at all providers (centralisation). Coupled with increasing costs and higher technology of medical care and efficiency drives putting pressure on public services, it is not surprising that networked models with different categories of provider are being advocated as the solution. In April 2013, the UK Department of Health launched a new system of networks across variety of clinical areas, neonatal care being one. Known as *Operational Delivery Networks* (ODNs), these 15 geographically defined networks were applied to existing networked specialisms with the potential for extension to other areas (NHS Commissioning Board, 2012). I chose to consider only the managed clinical network system because most of the data analysed were prior to the introduction of ODNs in April 2013, and even though they technically

came into being on this date, new systems take some time to embed. Despite managed clinical networks now being obsolete, the methods in this thesis will be applicable to the new ODNs for neonatal care and transferable to other clinical areas. Currently adult critical care, burns and major trauma are already delivered via networks in the UK NHS, and similar models are being considered for adult congenital heart disease, paediatric congenital cardiac surgery and paediatric neuroscience (NHS Commissioning Board, 2012).

9.6 Conclusion

This thesis has achieved the aim of developing methods to evaluate risk-adjusted mortality of preterm infants across networked NNUs in England using electronic patient records. In doing so, methodological contributions have been made regarding a novel way of specifying MM weightings, and new findings relating to the impact of misspecified weights on random effects. The thesis lays the foundations for future development to enable application of the methods to multiple births and for investigating challenges regarding the causal determinants of location of care. Practical implications could extend beyond neonatal medicine as centralisation, monitoring of clinical outcomes and electronic patient data become more integral to improving healthcare.

Bibliography

W. Aelvoet, N. Terryn, A. Blommaert, G. Molenberghs, N. Hens, F. De Smet, M. Callens, and P. Beutels. Community-acquired pneumonia (CAP) hospitalizations and deaths: is there a role for quality improvement through inter-hospital comparisons? *International journal for quality in health care : journal of the International Society for Quality in Health Care / ISQua*, 28(1):22–32, feb 2016. ISSN 1464-3677. doi: 10.1093/intqhc/mzv092.

M. Aitkin and N. Longford. Statistical Modelling Issues in School Effectiveness Studies. *Journal of the Royal Statistical Society. Series A (General)*, 149(1):1–43, 1986.

H. Akaike. A new look at the statistical model identification. *IEEE transactions on automatic control*, 19(6):716–723, 1974.

Z. Alfirevic. Tocolytics: do they actually work? *BMJ*, (345), 2012. doi: 10.1136/bmj.e6531.

M. Ali, Y. Jin, D. Kim, Z. De, J. Park, R. Ochiai, B. Dong, J. Clemens, and C. Acosta. Spatial risk for gender-specific adult mortality in an area of southern China. *International Journal of Health Geographics*, 6(31), 2007. doi: 10.1186/1476-072X-6-31.

C. V. Ananth, R. W. Platt, and D. A. Savitz. Regression models for clustered binary responses: implications of ignoring the intracluster correlation in an analysis of perinatal mortality in twin gestations. *Annals of epidemiology*, 15(4):293–301, apr 2005. ISSN 1047-2797. doi: 10.1016/j.annepidem.2004.08.007.

P.-Y. Ancel, F. Goffinet, P. Kuhn, B. Langer, J. Matis, et al. Survival and Morbidity of Preterm Children Born at 22 Through 34 Weeks' Gestation in France in 2011. *JAMA Pediatrics*, 169(3):230, mar 2015. ISSN 2168-6203. doi: 10.1001/jamapediatrics.2014.3351.

D. Annibale and R. Bissinger. The golden hour. *Advances in Neonatal Care*, 10(5):221–223, 2010.

C. C. Arnold, M. S. Kramer, C. A. Hobbs, F. H. McLean, and R. H. Usher. Very low birth weight: a problematic cohort for epidemiologic studies of very small or

immature neonates. *American journal of epidemiology*, 134(6):604–13, sep 1991. ISSN 0002-9262.

A. Ash, S. Fienberg, and T. Louis. Statistical issues in assessing hospital performance. Technical report, Commissioned by the Committee of Presidents of Statistical Societies for the Centers for Medicare and Medicaid Services, 2012. URL `http://escholarship.umassmed.edu/qhs{\_}pp/1114/`.

P. C. Austin. Bias in Penalized Quasi-Likelihood Estimation in Random Effects Logistic Regression Models When the Random Effects Are not Normally Distributed. *Communications in Statistics - Simulation and Computation*, 34(3):549–565, jul 2005. ISSN 0361-0918. doi: 10.1081/SAC-200068364.

P. C. Austin. Are (the log-odds of) hospital mortality rates normally distributed? Implications for studying variations in outcomes of medical care. *Journal of Evaluation in Clinical Practice*, 15(3):514–523, 2009. doi: 10.1111/j.1365-2753.2008.01053.x.

P. C. Austin, D. A. Alter, and J. V. Tu. The Use of Fixed-and Random-Effects Models for Classifying Hospitals as Mortality Outliers: A Monte Carlo Assessment. *Medical Decision Making*, 23(6):526–539, 2003. doi: 10.1177/0272989x03258443.

P. C. Austin, D. A. Alter, G. M. Anderson, and J. V. Tu. Impact of the choice of benchmark on the conclusions of hospital report cards. *American Heart Journal*, 148 (6):1041–1046, 2004. doi: 10.1016/j.ahj.2004.04.047.

P. Aylin and A. Bottle. Are hospital league tables calculated correctly? A commentary. *Public Health*, 121(12):905–906, dec 2007. ISSN 00333506. doi: 10.1016/j.puhe.2006.06.020.

P. Aylin, N. Best, A. Bottle, and C. Marshall. Following Shipman: a pilot system for monitoring mortality rates in primary care. *The Lancet*, 362(9382):485–491, 2003.

P. Aylin, A. Bottle, M.-H. Jen, and S. Middleton. HSMR mortality indicators. 2010. URL `https://www1.imperial.ac.uk/resources/3321CA24-A5BC-4A91-9CC9-12C74AA72FDC/`.

J. Bafumi and A. Gelman. Fitting multilevel models when predictors and group effects correlate. 2006. URL `http://papers.ssrn.com/sol3/papers.cfm?abstract{\_}id=1010095`.

S. Baldwin, D. Bauer, E. Stice, and P. Rohde. Evaluating models for partially clustered designs. *Psychological methods*, 16(2):149–165, 2011.

D. J. Barker. Fetal origins of coronary heart disease. *BMJ (Clinical research ed.)*, 311 (6998):171–4, jul 1995. ISSN 0959-8138.

O. Basso and A. J. Wilcox. Might Rare Factors Account for Most of the Mortality of Preterm Babies? *Epidemiology*, 22(3):320–327 10.1097/EDE.0b013e31821266c5, 2011.

C. Battersby, S. Santhakumaran, N. Modi, and K. Costeloe. Incidence And Outcomes Of Severe Necrotising Enterocolitis In Infants Less Than 32 Weeks Gestation: A Prospective Population Study. In *Archives of Disease in Childhood*, volume 99, pages A38–A38. BMJ Publishing Group Ltd and Royal College of Paediatrics and Child Health, oct 2014a. doi: 10.1136/archdischild-2014-307384.112. URL `http://adc.bmj.com/cgi/doi/10.1136/archdischild-2014-307384.112`.

C. Battersby, S. Santhakumaran, and M. Upton. The impact of a regional care bundle on maternal breast milk use in preterm infants: outcomes of the East of England quality improvement programme. *Archives of Disease in*, 2014b. doi: 10.1136/archdischild-2013-305475.

M. Bayes and M. Price. JSTOR: Philosophical Transactions (1683-1775), Vol. 53 (1763), pp. 370-418. *Philosophical Transactions (1683-1775)*, 1763.

G. Ben-Josef, L. S. Ott, S. B. Spivack, C. Wang, J. S. Ross, S. J. Shah, J. P. Curtis, N. Kim, H. M. Krumholz, and S. M. Bernheim. Payments for acute myocardial infarction episodes-of-care initiated at hospitals with and without interventional capabilities. *Circulation. Cardiovascular quality and outcomes*, 7(6):882–8, nov 2014. ISSN 1941-7705. doi: 10.1161/CIRCOUTCOMES.114.000927.

S. Binder, K. Hill, J. Meinzen-Derr, J. M. Greenberg, and V. Narendran. Increasing VLBW deliveries at subspecialty perinatal centers via perinatal outreach. *Pediatrics*, 127(3):487–93, mar 2011. ISSN 1098-4275. doi: 10.1542/peds.2010-1064.

S. M. Bird, C. Sir David, V. T. Farewell, G. Harvey, H. Tim, and S. Peter C. Performance indicators: good, bad, and ugly. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 168(1):1–27, 2005. doi: 10.1111/j.1467-985X.2004.00333.x.

H. Blencowe, S. Cousens, and M. Oestergaard. National, regional, and worldwide estimates of preterm birth rates in the year 2010 with time trends since 1990 for selected countries: a systematic analysis and. *The Lancet*, 379(9832):2162–2172, 2012.

B. Blondel, M. D. Kogan, G. R. Alexander, N. Dattani, M. S. Kramer, A. Macfarlane, and S. W. Wen. The Impact of the Increasing Number of Multiple Births on the Rates of Preterm Birth and Low Birthweight: An International Study. *American Journal of Public Health*, 92(8):1323–1330, aug 2002. ISSN 0090-0036. doi: 10.2105/AJPH.92.8.1323.

S. Bolisetty, N. Legge, B. Bajuk, and K. Lui. Preterm infant outcomes in New South Wales and the Australian Capital Territory. *Journal of paediatrics and child health*, 51(7):713–21, jul 2015. ISSN 1440-1754. doi: 10.1111/jpc.12848.

M. Bonet, B. Blondel, R. Agostino, and E. Combier. Variations in breastfeeding rates for very preterm infants between regions and neonatal units in Europe: results from the MOSAIC cohort. *Archives of Disease in*, (96):F450–F452, 2011. doi: 10.1136/adc.2009.179564.

A. Bottle, B. Jarman, and P. Aylin. Hospital Standardized Mortality Ratios: Sensitivity Analyses on the Impact of Coding. *Health Services Research*, 46(6pt1):1741–1761, dec 2011. ISSN 00179124. doi: 10.1111/j.1475-6773.2011.01295.x.

A. Bottle, R. Sanders, A. Mozid, and P. Aylin. Provider profiling models for acute coronary syndrome mortality using administrative data. *International journal of cardiology*, 168(1):338–343, 2013.

D. W. Bratzler, S.-L. T. Normand, Y. Wang, W. J. O'Donnell, M. Metersky, L. F. Han, M. T. Rapp, and H. M. Krumholz. An administrative claims model for profiling hospital 30-day mortality rates for pneumonia patients. *PloS one*, 6(4):e17401, 2011. ISSN 1932-6203. doi: 10.1371/journal.pone.0017401.

G. W. Brier. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1–3, jan 1950. ISSN 0027-0644. doi: 10.1175/1520-0493(1950)078⟨0001:VOFEIT⟩2.0.CO;2.

British Association of Perinatal Medicine. Service Standards for Hospitals Providing Neonatal Care (3rd edition). Technical report, 2010. URL `http://www.bapm.org/publications/documents/guidelines/BAPM{\_}Standards{\_}Final{\_}Aug2010.pdf`.

British Association of Perinatal Medicine. Categories of Care 2011. Technical report, 2011. URL `http://www.bapm.org/publications/documents/guidelines/CatsofcarereportAug11.pdf`.

M. Brookhart, D. Solomon, and P. Wang. Explained variation in a model of therapeutic decision making is partitioned across patient, physician, and clinic factors. *Journal of Clinical Epidemiology*, 59(1):18–25, 2006.

S. Brooks and A. Gelman. General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, 7(4):434–455, 1998.

W. J. Browne, H. Goldstein, and J. Rasbash. Multiple membership multiple classification (MMMC) models. *Statistical Modelling*, 1(2):103–124, 2001. doi: 10.1177/1471082x0100100202.

J. Bruckel, X. Liu, S. F. Hohmann, A. S. Karson, E. Mort, and D. M. Shahian. The denominator problem: national hospital quality measures for acute myocardial infarction. *BMJ quality & safety*, mar 2016. ISSN 2044-5423. doi: 10.1136/bmjqs-2015-004888.

P. Buekens and A. Wilcox. Why do small twins have a lower mortality rate than small singletons? *American Journal of Obstetrics and Gynecology*, 168(3):937–941, 1993.

J. F. Burgess, C. L. Christiansen, S. E. Michalak, and C. N. Morris. Medical profiling: improving standards and risk adjustments using hierarchical models. *Journal of Health Economics*, 19(3):291–309, 2000. doi: 10.1016/s0167-6296(99)00034-x.

A. Burton, D. G. Altman, P. Royston, and R. L. Holder. The design of simulation studies in medical statistics. *Statistics in Medicine*, 25(24):4279–4292, 2006. doi: 10.1002/sim.2673.

G. Cafri, D. Hedeker, and G. A. Aarons. An introduction and integration of cross-classified, multiple membership, and dynamic group random-effects models. *Psychological methods*, 20(4):407–21, dec 2015. ISSN 1939-1463. doi: 10.1037/met0000043.

S. Calling, X. Li, J. Sundquist, and K. Sundquist. Socioeconomic inequalities and infant mortality of 46,470 preterm infants born in Sweden between 1992 and 2006. *Paediatric and perinatal epidemiology*, 25(4):357–65, jul 2011. ISSN 1365-3016. doi: 10.1111/j.1365-3016.2011.01200.x.

M. J. Campbell, R. M. Jacques, J. Fotheringham, R. Maheswaran, and J. Nicholl. Developing a summary hospital mortality index: retrospective analysis in English hospitals over five years. *BMJ (Clinical research ed.)*, 344:e1001, 2012. ISSN 1756-1833.

B. Carlin and T. Louis. *Bayesian methods for data analysis*. Chapman & Hall, Boca Raton, Fla., 3rd edition, 2008.

J. B. Carlin, L. C. Gurrin, J. A. C. Sterne, R. Morley, and T. Dwyer. Regression models for twin studies: a critical review. *International Journal of Epidemiology*, 34 (5):1089–1099, oct 2005. doi: 10.1093/ije/dyi153.

T. Chandola, P. Clarke, R. D. Wiggins, and M. Bartley. Who you live with and where you live: setting the context for health using multiple membership multilevel models. *Journal of Epidemiology and Community Health*, 59(2):170–175, 2005. doi: 10.1136/jech.2003.019539.

Y. Chen, M. Rabilloud, and J. Thiolet. Benchmarking French regions according to their prevalence of healthcare-associated infections. *International journal for ...*, 25(5): 555–563, 2013.

Child Trends Databank. Preterm Births, 2015. URL http://www.childtrends.org/?indicators=preterm-births.

H. Chung and S. N. Beretvas. The impact of ignoring multiple membership data structures in multilevel models. *British Journal of Mathematical and Statistical Psychology*, 65(2):185–200, 2012. doi: 10.1111/j.2044-8317.2011.02023.x.

J. Cifuentes, J. Bronstein, C. S. Phibbs, R. H. Phibbs, S. K. Schmitt, and W. a. Carlo. Mortality in Low Birth Weight Infants According to Level of Neonatal Care at Hospital of Birth. *Pediatrics*, 109(5):745–751, may 2002. ISSN 0031-4005. doi: 10.1542/peds.109.5.745.

T. Cole, Y. Statnikov, S. Santhakumaran, and H. Pan. Birth weight and longitudinal growth in infants born below 32 weeks' gestation: a UK population study. *Archives of disease in*, 2013. doi: 10.1136/archdischild-2012-303536.

T. J. Cole, E. Hey, and S. Richmond. The PREM score: a graphical tool for predicting survival in very preterm births. *Archives of Disease in Childhood - Fetal and Neonatal Edition*, 95(1):F14–F19, 2010. doi: 10.1136/adc.2009.164533.

R. W. I. Cooke. Conventional birth weight standards obscure fetal growth restriction in preterm infants. *Archives of disease in childhood. Fetal and neonatal edition*, 92(3):F189–92, may 2007. ISSN 1359-2998. doi: 10.1136/adc.2005.089698.

K. Costeloe, E. Hennessy, A. T. Gibson, N. Marlow, A. R. Wilkinson, and f. t. E. S. Group. The EPICure Study: Outcomes to Discharge From Hospital for Infants Born at the Threshold of Viability. *Pediatrics*, 106(4):659–671, 2000.

K. L. Costeloe, E. M. Hennessy, S. Haider, F. Stacey, N. Marlow, and E. S. Draper. Short term outcomes after extreme preterm birth in England: comparison of two birth cohorts in 1995 and 2006 (the EPICure studies). *BMJ (Clinical research ed.)*, 345:e7976, jan 2012. ISSN 1756-1833.

D. R. Cox. Two Further Applications of a Model for Binary Regression. *Biometrika*, 45(3/4):562–565, 1958.

M. Croft, A. Read, N. de Klerk, and J. Hansen. Population based ascertainment of twins and their siblings, born in Western Australia 1980 to 1992, through the construction and validation of a maternally linked. *Twin*, 5(05):317–323, 2002.

J. M. Cusack, D. J. Field, and B. N. Manktelow. Impact of service changes on neonatal transfer patterns over 10 years. *Archives of Disease in Childhood - Fetal and Neonatal Edition*, 92(3):F181–F184, 2007. doi: 10.1136/adc.2006.105098.

M. Da Frè, A. Polo, D. Di Lallo, S. Piga, L. Gagliardi, et al. Size at birth by gestational age and hospital mortality in very preterm infants: results of the area-based ACTION

project. *Early human development*, 91(1):77–85, jan 2015. ISSN 1872-6232. doi: 10.1016/j.earlhumdev.2014.11.007.

M. J. Daniels and C. Gatsonis. Hierarchical Generalized Linear Models in the Analysis of Variations in Health Care Utilization. *Journal of the American Statistical Association*, 94(445):29–42, feb 1999.

I. Dégano, I. Subirana, M. Torre, M. Grau, and J. Vila. A European benchmarking system to evaluate in-hospital mortality rates in acute coronary syndrome: The EURHOBOP project. *International journal of*, 182:509–516, 2015.

V. Delaney-Black, L. O. Lubchenco, L. J. Butterfield, E. Goldson, B. L. Koops, and D. C. Lazotte. Outcome of very-low-birth-weight infants: are populations of neonates inherently different after antenatal versus neonatal referral? *Am J Obstet Gynecol*, 160(3):545–552, 1989.

E. R. DeLong, E. D. Peterson, D. M. DeLong, L. H. Muhlbaier, S. Hackett, and D. B. Mark. Comparing risk-adjustment methods for provider profiling. *Statistics in Medicine*, 16(23):2645–2664, 1997. doi: 10.1002/(sici)1097-0258(19971215)16:23⟨2645::aid-sim696⟩3.0.co;2-d.

Department for Communities and Local Government. The English Indices of Deprivation 2010. Technical report, Department of Communities and Local Government, London, 2011. URL `https://www.gov.uk/government/uploads/system/uploads/attachment{\_}data/file/6871/1871208.pdf`.

Department of Health. Neonatal intensive care services: report of the Department of Health Expert Working Group. Technical report, Department of Health, London, 2003.

Department of Health. Toolkit for high-quality neonatal services. Technical report, Department of Health, London, 2009.

Department of Health. Equity and excellence: Liberating the NHS. Technical report, Department of Health, London, 2010.

Department of Health. Birth ratios in England and Wales. Technical report, 2014. URL `https://www.gov.uk/government/uploads/system/uploads/attachment{\_}data/file/313559/Sex{\_}Selection{\_}in{\_}England{\_}and{\_}Wales{\_}analysis{\_}April{\_}2014{\_}NM{\_}comments.pdf`.

A. Diez-Roux. Multilevel analysis in public health research. *Annual review of public health*, 21:171–192, 2000.

J. B. Dimick, D. O. Staiger, and J. D. Birkmeyer. Ranking Hospitals on Surgical Mortality: The Importance of Reliability Adjustment. *Health Services Research*, 45 (6p1):1614–1629, 2010. doi: 10.1111/j.1475-6773.2010.01158.x.

L. Diya, E. Lesaffre, K. Van den Heede, W. Sermeus, and A. Vleugels. Establishing the relationship between nurse staffing and hospital mortality using a clustered discrete-time logistic model. *Statistics in Medicine*, 29(7-8):778–785, 2010. doi: 10.1002/sim.3756.

J. S. Dorling, D. J. Field, and B. Manktelow. Neonatal disease severity scoring systems. *Archives of Disease in Childhood - Fetal and Neonatal Edition*, 90(1):F11–F16, 2005. doi: 10.1136/adc.2003.048488.

L. Doyle and Victorian Infant Collaborative Study Group. Outcome at 5 years of age of children 23 to 27 weeks' gestation: refining the prognosis. *Pediatrics*, 108(1), 2001.

E. S. Draper, J. Zeitlin, A. C. Fenton, T. Weber, J. Gerrits, G. Martens, B. Misselwitz, and G. Breart. Investigating the variations in survival rates for very preterm infants in 10 European regions: the MOSAIC birth cohort. *Archives of Disease in Childhood - Fetal and Neonatal Edition*, 94(3):F158–F163, 2009. doi: 10.1136/adc.2008.141531.

S. Durrleman and R. Simon. Flexible regression models with cubic splines. *Statistics in medicine*, 8(5):551–561, 1989.

J. P. Empana, D. Subtil, and P. Truffert. In-hospital mortality of newborn infants born before 33 weeks of gestation depends on the initial level of neonatal care: the EPIPAGE study. *Acta Pædiatrica*, 92(3):346–351, 2003. doi: 10.1111/j.1651-2227.2003.tb00557.x.

EXPRESS Group, V. Fellman, L. Hellström-Westas, M. Norman, M. Westgren, K. Källén, H. Lagercrantz, K. Marsál, F. Serenius, and M. Wennergren. One-year survival of extremely preterm infants after active perinatal care in Sweden. *JAMA*, 301(21):2225–33, jun 2009. ISSN 1538-3598. doi: 10.1001/jama.2009.771.

A. C. Fenton, A. Leslie, and C. H. Skeoch. Optimising neonatal transfer. *Archives of Disease in Childhood - Fetal and Neonatal Edition*, 89(3):F215–F219, 2004. doi: 10.1136/adc.2002.019711.

A. Fielding and H. Goldstein. Cross-classified and Multiple Membership Structures in Multilevel Models: An Introduction and Review. Technical report, Department for Education and Skills Publications, Nottingham, 2006.

A. Fielding and M. Yang. Generalized linear mixed models for ordered responses in complex multilevel structures: effects beneath the school or college in education. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 168(1):159–183, 2005. doi: 10.1111/j.1467-985X.2004.00342.x.

S. A. Friedman, E. Schiff, L. Kao, J. Kuint, and B. M. Sibai. Do twins mature earlier than singletons? Results from a matched cohort study. *American Journal of Obstetrics and Gynecology*, 176(6):1193–1199, 1997. ISSN 00029378. doi: 10.1016/S0002-9378(97)70334-X.

C. Gale, A. Hay, C. Philipp, R. Khan, S. Santhakumaran, and N. Ratnavel. In-utero transfer is too difficult: results from a prospective study. *Early human development*, 88(3):147–50, mar 2012a. ISSN 1872-6232. doi: 10.1016/j.earlhumdev.2011.07.016.

C. Gale, S. Santhakumaran, S. Nagarajan, Y. Statnikov, and N. Modi. Impact of managed clinical networks on NHS specialist neonatal services in England: population based study. *BMJ*, 344:e2105:doi: 10.1136/bmj.e2105, 2012b.

C. Gale, N. Modi, o. b. o. t. W. trial development Group, C. Gale, N. Modi, et al. Neonatal randomised point-of-care trials are feasible and acceptable in the UK: results from two national surveys. *Archives of Disease in Childhood - Fetal and Neonatal Edition*, 101(1):86–87, jan 2016a. ISSN 1359-2998. doi: 10.1136/archdischild-2015-308882.

C. Gale, I. Morris, and S. Board. The UK National Neonatal Research Database: using neonatal data for research, quality improvement and more. *Archives of Disease in Childhood Education & Practice*, 2016b. doi: 10.1136/archdischild-2015-309928.

M. O. Gardner, R. L. Goldenberg, S. P. Cliver, J. M. Tucker, K. G. Nelson, and R. L. Copper. The origin and outcome of preterm twin pregnancies. *Obstetrics and gynecology*, 85(4):553–7, apr 1995. ISSN 0029-7844. doi: 10.1016/0029-7844(94)00455-M.

C. Gatsonis, S. Normand, C. Liu, and C. Morris. Geographic variation of procedure utilization: a hierarchical model approach. *Medical Care*, 31(5):YS54–YS59, 1993.

C. Gatsonis, A. Epstein, J. Newhouse, and S. Normand. Variations in the utilization of coronary angiography for elderly patients with an acute myocardial infarction: an analysis using hierarchical logistic regression. *Medical care*, 33(6):625–642, 1995.

A. Gelfand and A. Smith. Sampling-based approaches to calculating marginal densities. *Journal of the American statistical*, 85(410):398–409, 1990.

A. Gelman. Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, 1(3):515–533, 2006.

A. Gelman and D. B. Rubin. Inference from iterative simulation using multiple sequences. *Statistical Science*, 7(4):457–472, 1992.

A. Gelman, X. Meng, and H. Stern. Posterior predictive assessment of model fitness via realized discrepancies. *Statistica sinica*, 6(4):733–760, 1996.

A. Gelman, Y. Goegebeur, F. Tuerlinckx, and I. V. Mechelen. Diagnostic checks for discrete data regression models using posterior predictive simulations. *Applied Statistics*, 49(2):247–268, 2000.

A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian data analysis*. Taylor & Francis, Florida, 2nd edition, 2004.

S. Geman and D. Geman. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on pattern analysis and machine intelligence*, 6:721–741, 1984.

J. Geweke. Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. Technical report, Federal Reserve Bank of Minneapolis and University of Minnesota, Minneapolis, 1991. URL `https://minneapolisfed.org/research/SR/SR148.pdf`.

W. R. Gilks, S. Richardson, and D. J. Spiegelhalter. *Markov chain Monte Carlo in practice*. Chapman & Hall, London, 1996. ISBN 0412055511.

A. B. Gill, L. Bottomley, S. Chatfield, and C. Wood. Perinatal transport: problems in neonatal intensive care capacity. *Archives of Disease in Childhood - Fetal and Neonatal Edition*, 89(3):F220–F223, 2004. doi: 10.1136/adc.2003.028159.

M. Gissler and A. Mohangoo. Perinatal health monitoring in Europe: results from the EURO-PERISTAT project. *Informatics for Health and Social Care*, 35(2):64–79, 2010.

L. Glance, A. Dick, T. Osler, and D. Mukamel. Using hierarchical modeling to measure ICU quality. *Intensive Care Medicine*, 29(12):2223–2229, 2003.

L. G. Glance, A. Dick, T. M. Osler, Y. Li, and D. B. Mukamel. Impact of Changing the Statistical Methodology on Hospital and Surgeon Ranking: The Case of the New York State Cardiac Surgery Report Card. *Medical Care*, 44(4):311–319 10.1097/01.mlr.0000204106.64619.2a, 2006.

L. G. Glance, T. M. Osler, D. B. Mukamel, W. Meredith, and A. W. Dick. Impact of Statistical Approaches for Handling Missing Data on Trauma Center Quality. *Annals of Surgery*, 249(1):143–148 10.1097/SLA.0b013e31818e544b, 2009.

R. Goldenberg and M. Gravett. The preterm birth syndrome: issues to consider in creating a classification system. *American Journal of Obstetrics and Gynecology*, 206 (2):113–118, 2012.

R. L. Goldenberg, J. F. Culhane, J. D. Iams, and R. Romero. Epidemiology and causes of preterm birth. *The Lancet*, 371(9606):75–84, 2008. doi: 10.1016/s0140-6736(08) 60074-4.

H. Goldstein. *Multilevel Statistical Models*. Hodder Arnold, London, 3rd edition, 2002. ISBN 0340806559.

H. Goldstein and D. J. Spiegelhalter. League Tables and Their Limitations: Statistical Issues in Comparisons of Institutional Performance. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 159(3):385–443, 1996.

H. Goldstein, W. Browne, and J. Rasbash. Partitioning Variation in Multilevel Models. *Understanding Statistics*, 1(4):223–231, dec 2002. ISSN 1534-844X.

H. Goldstein, S. Burgess, and B. McConnell. Modelling the Impact of Pupil Mobility on School Differences in Educational Achievement. 2006.

H. Goldstein, S. Burgess, and B. McConnell. Modelling the effect of pupil mobility on school differences in educational achievement. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 170(4):941–954, 2007. doi: 10.1111/j.1467-985X.2007.00491.x.

D. Gomez, A. Alali, W. Xiong, B. Zarzaur, and N. Mann. Definitive care in level 3 trauma centres after severe injury: A comparison of clinical characteristics and outcomes. *Injury*, 46(9):1790–1795, 2015.

L. Gortner, B. Misselwitz, D. Milligan, J. Zeitlin, L. Kollée, et al. Rates of Bronchopulmonary Dysplasia in Very Preterm Neonates in Europe: Results from the MOSAIC Cohort. *Neonatology*, 99(2):112–117, 2011. doi: 10.1159/000313024.

R. Gray and M. McCormick. Socioeconomic inequalities in survival in neonates. *BMJ*, (229):b5041, 2009. doi: 10.1136/bmj.b5041.

J. Green and N. Wintfeld. Report cards on cardiac surgeons. Assessing New York State's approach. *The New England journal of medicine*, 332(18):1229–32, may 1995. ISSN 0028-4793. doi: 10.1056/NEJM199505043321812.

K. Guenther, W. Vach, W. Kachel, I. Bruder, and R. Hentschel. Auditing Neonatal Intensive Care: Is PREM a Good Alternative to CRIB for Mortality Risk Adjustment in Premature Infants? *Neonatology*, 108(3):172–8, jan 2015. ISSN 1661-7819. doi: 10.1159/000433414.

H. L. Halliday. Surfactants: past, present and future. *Journal of perinatology : official journal of the California Perinatal Association*, 28 Suppl 1(S1):S47–56, may 2008. ISSN 0743-8346. doi: 10.1038/jp.2008.50.

F. E. Harrell. *Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis*. Springer-Verlag, New York, 2001.

F. E. Harrell, K. L. Lee, and D. B. Mark. Multivariable Prognostic Models: Issues In Developing Models, Evaluating Assumptions And Adequacy, And Measuring And Reducing Errors. *Statistics in Medicine*, 15(4):361–387, 1996. doi: 10.1002/(sici)1097-0258(19960229)15:4⟨361::aid-sim168⟩3.0.co;2-4.

S. Hassani, A. Lindman, D. Kristoffersen, and O. Tomic. 30-Day Survival Probabilities as a Quality Indicator for Norwegian Hospitals: Data Management and Analysis. *PloS one*, 2015. doi: 10.1371/journal.pone.0136547.

T. Hastie and R. Tibshirani. Generalized additive models. *Statistical science*, 1(3): 297–310, 1986.

Health and Social Care Information Centre. Statistics on Smoking, England 2016, 2016a. URL `http://www.hscic.gov.uk/article/2021/Website-Search?productid=21003{\&}q=smoking{\&}sort=Relevance{\&}size=10{\&}page=1{\&}area=both{\#}top`.

Health and Social Care Information Centre. Summary Hospital-level Mortality Indicator (SHMI) - Deaths associated with hospitalisation, England, January 2015-December 2015, Quarterly report. Technical report, 2016b. URL `http://www.hscic.gov.uk/catalogue/PUB20949/shmi-deat-hosp-eng-jan-15-dec-15-rep.pdf`.

Health and Social Care Information Centre Clinical Indicators Team. Indicator Specification: Summary Hospital-level Mortality Indicator version 1.22. Technical report, 2016. URL `http://www.hscic.gov.uk/SHMI`.

P. Heidelberger and P. Welch. Simulation run length control in the presence of an initial transient. *Operations Research*, 1983. doi: 10.1287/opre.31.6.1109.

M. A. Hernán, S. Hernández-Díaz, M. M. Werler, and A. A. Mitchell. Causal Knowledge as a Prerequisite for Confounding Evaluation: An Application to Birth Defects Epidemiology. *American Journal of Epidemiology*, 155(2):176–184, 2002. doi: 10.1093/aje/155.2.176.

P. Hill and H. Goldstein. Multilevel modeling of educational data with cross-classification and missing identification for units. *Journal of Educational and Behavioral Statistics*, 23(2):117–128, 1998.

E. Himpens, C. Van den Broeck, A. Oostra, P. Calders, and P. Vanhaesebrouck. Prevalence, type, distribution, and severity of cerebral palsy in relation to gestational age: a meta-analytic review. *Developmental medicine and child neurology*, 50(5):334–40, may 2008. ISSN 0012-1622. doi: 10.1111/j.1469-8749.2008.02047.x.

T. P. Hofer, R. A. Hayward, S. Greenfield, E. H. Wagner, S. H. Kaplan, and W. G. Manning. The Unreliability of Individual Physician "Report Cards" for Assessing the

Costs and Quality of Care of a Chronic Disease. *JAMA: The Journal of the American Medical Association*, 281(22):2098–2105, 1999. doi: 10.1001/jama.281.22.2098.

M. Hohlagschwandtner, P. Husslein, K. Klebermass, M. Weninger, A. Nardi, and M. Langer. Perinatal mortality and morbidity. *Archives of Gynecology and Obstetrics*, 265(3):113–118, 2001. doi: 10.1007/s004040100197.

J. D. Horbar, J. H. Carpenter, G. J. Badger, M. J. Kenny, R. F. Soll, K. A. Morrow, and J. S. Buzas. Mortality and neonatal morbidity among infants 501 to 1500 grams from 2000 to 2009. *Pediatrics*, 129(6):1019–26, jun 2012. ISSN 1098-4275. doi: 10.1542/peds.2011-3028.

D. Hosmer, S. Lemeshow, and R. Sturdivant. *Applied logistic regression*. 2013.

L. I. Iezzoni. Using risk-adjusted outcomes to assess clinical practice: An overview of issues pertaining to risk adjustment. *The Annals of Thoracic Surgery*, 58(6):1822–1826, 1994. doi: 10.1016/0003-4975(94)91721-3.

L. I. Iezzoni. *Risk Adjustment for Measuring Health Care Outcomes*. Health Administration Press, 3rd edition, 2003.

I. Ingemarsson. Gender aspects of preterm birth. *BJOG : an international journal of obstetrics and gynaecology*, 110 Suppl:34–8, may 2003. ISSN 1470-0328.

M. Jakobsson, M. Gissler, J. Paavonen, and A.-M. Tapper. The incidence of preterm deliveries decreases in Finland. *BJOG : an international journal of obstetrics and gynaecology*, 115(1):38–43, jan 2008. ISSN 1471-0528. doi: 10.1111/j.1471-0528.2007.01565.x.

S. Jan, G. Slap, K. Smith-Whitley, D. Dai, R. Keren, and D. M. Rubin. Association of hospital and provider types on sickle cell disease outcomes. *Pediatrics*, 132(5):854–61, nov 2013. ISSN 1098-4275. doi: 10.1542/peds.2013-0089.

J. Jollis and P. Romano. Pennsylvania's focus on heart attack—grading the scorecard. *New England Journal of Medicine*, 338:983–987, 1998.

H. E. Jones, D. I. Ohlssen, and D. J. Spiegelhalter. Use of the false discovery rate when comparing multiple health care providers. *Journal of Clinical Epidemiology*, 61(3):232–240.e2, 2008. doi: 10.1016/j.jclinepi.2007.04.017.

K. S. Joseph, S. Liu, J. Rouleau, S. Lisonkova, J. A. Hutcheon, R. Sauve, A. C. Allen, and M. S. Kramer. Influence of definition based versus pragmatic birth registration on international comparisons of perinatal and infant mortality: population based retrospective study. *BMJ*, 344, 2012. doi: 10.1136/bmj.e746.

S. T. Kempley, Y. Baki, G. Hayter, N. Ratnavel, E. Cavazzoni, and T. Reyes. Effect of a centralised transfer service on characteristics of inter-hospital neonatal transfers. *Archives of Disease in Childhood - Fetal and Neonatal Edition*, 2006. doi: 10.1136/adc.2006.106047.

B. Keogh. Review into the quality of care and treatment provided by 14 hospital trusts in England: overview report. Technical report, NHS England, London, 2013. URL `http://www.nhs.uk/NHSEngland/bruce-keogh-review/Documents/outcomes/keogh-review-final-report.pdf`.

J. L. Kiely. What is the population-based risk of preterm birth among twins and other multiples? *Clinical obstetrics and gynecology*, 41(1):3–11, mar 1998. ISSN 0009-9201.

N. Kim, S. M. Bernheim, L. S. Ott, L. Han, S. B. Spivack, X. Xu, M. Volpe, A. Liu, and H. M. Krumholz. An administrative claims measure of payments made for Medicare patients for a 30-day episode of care for acute myocardial infarction. *Medical care*, 53 (6):542–9, jun 2015. ISSN 1537-1948. doi: 10.1097/MLR.0000000000000361.

L. A. A. Kollée, P. P. Verloove-Vanhorick, R. A. Verwey, R. Brand, and J. H. Ruys. Maternal and Neonatal Transport: Results of a National Collaborative Survey of Preterm and Very Low Birth Weight Infants in the Netherlands. *Obstetrics & Gynecology*, 72 (5):729–732, 1988.

E. W. Kopf. Florence Nightingale as Statistician. *Publications of the American Statistical Association*, 15(116):388, dec 1916. ISSN 15225437. doi: 10.2307/2965763.

M. Kosseim, N. E. Mayo, S. Scott, J. A. Hanley, J. Brophy, B. Gagnon, and L. Pilote. Ranking hospitals according to acute myocardial infarction mortality: should transfers be included? *Medical Care*, 44(7):664–670, 2006.

M. Kramer, L. Goulet, and J. Lydon. Socioeconomic disparities in preterm birth: causal pathways and mechanisms. *Paediatric and Perinatal Epidemiology*, 15(s2):104–123, 2001.

D. T. Kristoffersen, J. Helgeland, J. Clench-Aas, P. Laake, M. B. Veierød, et al. Comparing hospital mortality – how to count does matter for patients hospitalized for acute myocardial infarction (AMI), stroke and hip fracture. *BMC Health Services Research*, 12(1):364, dec 2012. ISSN 1472-6963. doi: 10.1186/1472-6963-12-364.

H. Krumholz, S. Normand, J. Spertus, and D. Shahian. Measuring performance for treating heart attacks and heart failure: the case for outcomes measurement. *Health Affairs*, 26(1):75–85, 2007.

H. M. Krumholz, R. G. Brindis, J. E. Brush, D. J. Cohen, A. J. Epstein, et al. Standards for Statistical Models Used for Public Reporting of Health Outcomes. *Circulation*, 113(3):456–462, 2006. doi: 10.1161/circulationaha.105.170769.

J. J. Kurinczuk, J. Hollowell, P. Boyd, L. Oakley, P. Brocklehurst, and R. Gray. Inequalities in infant mortality project briefing paper 4. The contribution of congenital anomalies to infant mortality. Technical report, National Perinatal Epidemiology Unit, Oxford, 2010. URL `https://www.npeu.ox.ac.uk/downloads/files/infant-mortality/Infant-Mortality-Briefing-Paper-4.pdf`.

S. Kusuda, M. Fujimura, A. Uchiyama, and S. Totsu. Trends in morbidity and mortality among very-low-birth-weight infants from 2003 to 2008 in Japan. *Pediatric Research*, 75:531–538, 2012.

P. Lambert, A. Sutton, and P. Burton. How vague is vague? A simulation study of the impact of the use of vague prior distributions in MCMC using WinBUGS. *Statistics in*, 2005.

R. F. Lamont, P. D. Dunlop, P. Crowley, M. I. Levene, and M. G. Elder. Comparative mortality and morbidity of infants transferred in utero or postnatally. *J Perinat Med*, 11(4):200–203, 1983.

S. M. Lasswell, W. D. Barfield, R. W. Rochat, and L. Blackmon. Perinatal regionalization for very low-birth-weight and very preterm infants: a meta-analysis. *JAMA*, 304 (9):992–1000, sep 2010. ISSN 1538-3598. doi: 10.1001/jama.2010.1226.

J. E. Lawn, S. Cousens, and J. Zupan. 4 million neonatal deaths: when? Where? Why? *Lancet*, 365(9462):891–900, jan 2005. ISSN 1474-547X. doi: 10.1016/S0140-6736(05)71048-5.

G. Leckie. The complexity of school and neighbourhood effects and movements of pupils on school differences in models of educational achievement. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 172(3):537–554, 2009. doi: 10.1111/j.1467-985X.2008.00577.x.

H. C. Lee, M. V. Bennett, J. Schulman, and J. B. Gould. Accounting for variation in length of NICU stay for extremely low birth weight infants. *Journal of perinatology : official journal of the California Perinatal Association*, 33(11):872–6, nov 2013. ISSN 1476-5543. doi: 10.1038/jp.2013.92.

K. J. Lee and S. G. Thompson. Flexible parametric models for random-effects distributions. *Statistics in Medicine*, 27(3):418–434, 2008. doi: 10.1002/sim.2897.

K. Leveno, J. Quirk, P. Whalley, and W. Herbert. Fetal lung maturation in twin gestation. *American journal of*, 148(4):405–411, 1984.

A. Leyland and H. Goldstein. *Multilevel Modelling of Health Statistics*. Wiley, Chichester, 2001.

A. Leyland and Ø. Næss. The effect of area of residence over the life course on subsequent mortality. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 172(3):555–578, 2009.

R. Lilford and P. Pronovost. Using hospital mortality rates to judge hospital performance: a bad idea that just won't go away. *BMJ (Clinical research ed.)*, 340(May): c2016, jan 2010. ISSN 1756-1833.

R. J. A. Little. *Statistical analysis with missing data*. Wiley, New York ; Chichester, 1987. ISBN 0471802549.

D. Lunn, C. Jackson, N. Best, A. Thomas, and D. Spiegelhalter. *The BUGS book: A practical introduction to Bayesian analysis*. CRC press, 2012.

M. F. MacDorman, T. Mathews, A. D. Mohangoo, and J. Zeitlin. International Comparisons of Infant Mortality and Related Factors: United States and Europe, 2010. Technical report, 2014. URL `http://www.cdc.gov/nchs/data/nvsr/nvsr63/nvsr63{\_}05.pdf`.

Y. MacNab, Z. Qiu, P. Gustafson, C. B. Dean, A. Ohlsson, and S. K. Lee. Hierarchical Bayes Analysis of Multilevel Health Services Data: A Canadian Neonatal Mortality Study. *Health Services and Outcomes Research Methodology*, 5(1):5–26, jul 2005. ISSN 1387-3741. doi: 10.1007/s10742-005-5561-1.

S. Manda and R. Meyer. Bayesian inference for recurrent events data using time dependent frailty. *Statistics in medicine*, 24(8):1263–1274, 2005.

B. N. Manktelow, E. S. Draper, and D. J. Field. Predicting neonatal mortality among very preterm infants: a comparison of three versions of the CRIB score. *Archives of Disease in Childhood - Fetal and Neonatal Edition*, 95(1):F9–F13, 2010. doi: 10.1136/adc.2008.148015.

B. N. Manktelow, S. Seaton, D. Field, and E. Draper. Population based estimates of in-unit survival for very preterm babies. *Pediatrics*, 131(2):e425–32, 2013.

B. N. Manktelow, L. K. Smith, S. E. Seaton, P. Hyman-Taylor, J. J. Kurinczuk, D. J. Field, and E. S. Draper. MBRRACE-UK Perinatal Mortality Surveillance Report. Technical report, The Infant Mortality and Morbidity Studies, Department of Health Sciences, University of Leicester, Leicester, 2016. URL `https://www.npeu.ox.ac.uk/mbrrace-uk/reports`.

C. Mann, C. Ward, M. Grubb, B. Hayes-Gill, J. Crowe, N. Marlow, and D. Sharkey. Marked variation in newborn resuscitation practice: a national survey in the UK. *Resuscitation*, 83(5):607–11, may 2012. ISSN 1873-1570. doi: 10.1016/j.resuscitation.2012.01.002.

J. P. Marcin and P. S. Romano. Impact of between-hospital volume and within-hospital volume on mortality and readmission rates for trauma patients in California. *Critical care medicine*, 32(7):1477–83, jul 2004. ISSN 0090-3493.

N. Marlow, C. Bennett, E. S. Draper, E. M. Hennessy, A. S. Morgan, and K. L. Costeloe. Perinatal outcomes for extremely preterm babies in relation to place of birth in England: the EPICure 2 study. *Archives of disease in childhood. Fetal and neonatal edition*, 99(3):F181–8, may 2014. ISSN 1468-2052. doi: 10.1136/archdischild-2013-305555.

E. C. Marshall, D. J. Spiegelhalter, C. Sanderson, and M. McKee. Reliability of league tables of in vitro fertilisation clinics: retrospective analysis of live birth rates. *BMJ*, 316(7146):1701–1705, 1998.

L. Marston, J. L. Peacock, K. Yu, P. Brocklehurst, S. A. Calvert, A. Greenough, and N. Marlow. Comparing methods of analysing datasets with small clusters: case studies using four paediatric datasets. *Paediatric and perinatal epidemiology*, 23(4):380–92, jul 2009. ISSN 1365-3016. doi: 10.1111/j.1365-3016.2009.01046.x.

A. Mason, N. Best, S. Richardson, and I. Plewis. Strategy for modelling non-random missing data mechanisms in observational studies using Bayesian methods. Technical report, Imperial College London, London, 2010.

M. C. McCormick. The Contribution of Low Birth Weight to Infant Mortality and Childhood Morbidity. *New England Journal of Medicine*, 312(2):82–90, 1985. doi: doi:10.1056/NEJM198501103120204.

S. Medlock, A. C. J. Ravelli, P. Tamminga, B. W. M. Mol, and A. Abu-Hanna. Prediction of Mortality in Very Premature Infants: A Systematic Review of Prediction Models. *PLoS ONE*, 6(9):e23441, 2011. doi: 10.1371/journal.pone.0023441.

M. E. Miller, S. L. Hui, and W. M. Tierney. Validation techniques for logistic regression models. *Statistics in Medicine*, 10(8):1213–1226, 1991. doi: 10.1002/sim.4780100805.

H. Minakami and I. Sato. Reestimating date of delivery in multifetal pregnancies. *JAMA*, 275(18):1432–1434, 1996.

M. Mizrahi, B. Furman, I. Shoham-Vardi, H. Vardi, E. Maymon, and M. Mazor. Perinatal outcome and peripartum complications in preterm singleton and twins deliveries: a comparative study. *European journal of obstetrics, gynecology, and reproductive biology*, 87(1):55–61, nov 1999. ISSN 0301-2115.

H. D. Modanlou, W. L. Dorchester, A. Thorosian, and R. K. Freeman. Antenatal Versus Neonatal Transport to a Regional Perinatal Center: A Comparison Between Matched Pairs. *Obstetrics & Gynecology*, 53(6):725–729, 1979.

N. Modi, J. Abbot, D. Ashby, P. Brocklehurst, Z. Chivers, et al. Medicines for Neonates: using routinely recorded clinical data for research. 2016.

M. A. Mohamed, A. Nada, and H. Aly. Day-by-Day Postnatal Survival in Very Low Birth Weight Infants. *Pediatrics*, 126(2):e360–e366, 2010. doi: 10.1542/peds.2009-2810.

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

M. A. Mohammed, B. N. Manktelow, and T. P. Hofer. Comparison of four methods for deriving hospital standardised mortality ratios from a single hierarchical logistic regression model. *Statistical methods in medical research*, pages 0962280212465165–, dec 2012. ISSN 1477-0334. doi: 10.1177/0962280212465165.

Y. Mundlak. On the pooling of time series and cross section data. *Econometrica: journal of the Econometric Society*, 1978.

D. Murphy, P. Hope, and A. Johnson. Neonatal risk factors for cerebral palsy in very preterm babies: case-control study. *BMJ*, 314, 1997. doi: 10.1136/bmj.314.7078.404.

National Audit Office. Caring for Vulnerable Babies: The reorganisation of neonatal services in England. 2007.

National Institute for Child Health and Human Development. NICHD Neonatal Research Network:Extremely Preterm Birth Outcome Data. URL `https://www.nichd.nih.gov/about/org/der/branches/ppb/programs/epbo/pages/epbo{\_}case.aspx`.

National Institute for Health and Care Excellence. Antenatal care for uncomplicated pregnancies (CG62). Technical report, 2008. URL `https://www.nice.org.uk/guidance/cg62/resources/antenatal-care-for-uncomplicated-pregnancies-975564597445`.

National Quality Forum. Risk adjustment for socioeconomic status or other sociodemographic factors. Technical report, Washington, DC, 2014.

Neonatal Critical Care Clinical Reference Group. Neonatal Critical Care Service Specification E08/S/a. Technical report, NHS England, 2015. URL `https://www.england.nhs.uk/commissioning/wp-content/uploads/sites/12/2015/01/e08-serv-spec-neonatal-critical.pdf`.

Neonatal Data Analysis Unit. Analysis of neonatal mortality data 2013-2014. Technical report, Neonatal Data Analysis Unit, Imperial College London, London, 2015. URL `https://www1.imperial.ac.uk/neonataldataanalysis/supporting/na/`.

Neonatal Data Analysis Unit. Neonatal Data Analysis Unit, 2016. URL https://www1.imperial.ac.uk/neonataldataanalysis/.

M. T. Neto. Perinatal care in Portugal: effects of 15 years of a regionalized system. *Acta paediatrica (Oslo, Norway : 1992)*, 95(11):1349–52, nov 2006. ISSN 0803-5253. doi: 10.1080/08035250600615135.

NHS Choices. NHSChoices Home Page. URL http://www.nhs.uk/pages/home.aspx.

NHS Commissioning Board. The Way Forward: Operational Delivery Networks. Technical report, 2012. URL https://www.england.nhs.uk/wp-content/uploads/2012/12/develop-odns.pdf.

J. Nicholl. Case-mix adjustment in non-randomised observational evaluations: the constant risk fallacy. *Journal of epidemiology and community health*, 61(11):1010–3, nov 2007. ISSN 0143-005X. doi: 10.1136/jech.2007.061747.

F. Nightingale. *Notes on hospitals*. 1863.

S.-L. T. Normand and D. M. Shahian. Statistical and Clinical Aspects of Hospital Outcomes Profiling. *Statistical Science*, 22(2):206–226, 2007.

S.-L. T. Normand, M. E. Glickman, and C. A. Gatsonis. Statistical Methods for Profiling Providers of Medical Care: Issues and Applications. *Journal of the American Statistical Association*, 92(439):803–814, 1997.

Nuffield Council on Bioethics. Critical care decisions in fetal and neonatal medicine: ethical issues. Technical report, 2006. URL http://nuffieldbioethics.org/project/neonatal-medicine/.

C. P. F. O'Donnell, C. O. F. Kamlin, P. G. Davis, J. B. Carlin, and C. J. Morley. Interobserver variability of the 5-minute Apgar score. *The Journal of pediatrics*, 149 (4):486–9, oct 2006. ISSN 0022-3476. doi: 10.1016/j.jpeds.2006.05.040.

Office of National Statistics. Focus on Gender. Technical report, 2008. URL http://webarchive.nationalarchives.gov.uk/20160105160709/http://www.ons.gov.uk/ons/publications/re-reference-tables.html?edition=tcm{\%}253A77-51144.

Office of National Statistics. Gestation-specific infant mortality in England and Wales 2013. Technical report, Office of National Statistics, Newport, 2014. URL http://web.ons.gov.uk/ons/publications/re-reference-tables.html?edition=tcm:77-395938.

Office of National Statistics. Birth Characteristics in England and Wales 2014, 2015. URL http://www.ons.gov.uk/ons/rel/vsob1/birth-characteristics-in-england-and-wales/2014/index.html.

G. Ohel, M. Granat, D. Zeevi, A. Golan, and S. Wexler. Advanced ultrasonic placental maturation in twin pregnancies. *American journal of*, 1987.

D. I. Ohlssen, L. D. Sharples, and D. J. Spiegelhalter. Flexible random-effects models using Bayesian semi-parametric models: applications to institutional comparisons. *Stat Med*, 26(9):2088–2112, 2007a. doi: 10.1002/sim.2666.

D. I. Ohlssen, L. D. Sharples, and D. J. Spiegelhalter. A hierarchical modelling framework for identifying unusual performance in health care providers. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 170(4):865–890, oct 2007b. ISSN 0964-1998. doi: 10.1111/j.1467-985X.2007.00487.x.

E. Oken, K. P. Kleinman, J. Rich-Edwards, M. W. Gillman, M. McCormick, et al. A nearly continuous measure of birth weight for gestational age using a United States national reference. *BMC Pediatrics*, 3(1):6, 2003. ISSN 14712431. doi: 10.1186/1471-2431-3-6.

OpenBUGS Project Management Group. OpenBUGS, 2014.

E. Oral, A. Cağdaş, A. Gezer, S. Kaleli, K. Aydinli, and F. Öçer. Perinatal and maternal outcomes of fetal macrosomia. *European Journal of Obstetrics & Gynecology and Reproductive Biology*, 99(2):167–171, dec 2001. ISSN 03012115. doi: 10.1016/S0301-2115(01)00416-X.

G. Parry, J. Tucker, and W. Tarnow-Mordi. CRIB II: an update of the clinical risk index for babies score. *The Lancet*, 361(9371):1789–1791, 2003. doi: 10.1016/s0140-6736(03)13397-1.

R. Patel, S. Kandefer, and M. Walsh. Causes and timing of death in extremely premature infants from 2000 through 2011. *New England Journal of Medicine*, 372:331–340, 2015.

J. L. Peacock, L. Marston, N. Marlow, S. A. Calvert, and A. Greenough. Neonatal and infant outcome in boys and girls born very prematurely. *Pediatric research*, 71(3):305–10, mar 2012. ISSN 1530-0447. doi: 10.1038/pr.2011.50.

C. S. Phibbs. Managed clinical networks in neonatal care. *BMJ (Clinical research ed.)*, 344(apr03_1):e2423, jan 2012. ISSN 1756-1833. doi: 10.1136/bmj.e2423.

C. S. Phibbs, J. M. Bronstein, E. Buxton, and R. H. Phibbs. The effects of patient volume and level of care at the hospital of birth on neonatal mortality. *JAMA : the journal of the American Medical Association*, 276(13):1054–9, oct 1996. ISSN 0098-7484.

C. S. Phibbs, L. C. Baker, A. B. Caughey, B. Danielsen, S. K. Schmitt, and R. H. Phibbs. Level and volume of neonatal intensive care and mortality in very-low-birth-

weight infants. *The New England journal of medicine*, 356(21):2165–75, may 2007. ISSN 1533-4406. doi: 10.1056/NEJMsa065029.

D. Pitches, M. Mohammed, and R. Lilford. What is the empirical evidence that hospitals with higher-risk adjusted mortality rates provide poorer quality care? A systematic review of the literature. *BMC Health Services Research*, 7(1):91, 2007.

M. Plummer, N. Best, K. Cowles, and K. Vines. CODA: Convergence Diagnosis and Output Analysis for MCMC, 2006.

M. M. Pollack, M. A. Koch, D. A. Bartel, I. Rapoport, R. Dhanireddy, A. A. E. El-Mohandes, K. Harkavy, and K. N. S. Subramanian. A Comparison of Neonatal Mortality Risk Prediction Models in Very Low Birth Weight Infants. *PEDIATRICS*, 105 (5):1051–1057, may 2000. ISSN 0031-4005. doi: 10.1542/peds.105.5.1051.

D. Pregibon. Logistic regression diagnostics. *The Annals of Statistics*, 9(4):705–724, 1981.

R Core Team. R: A Language and Environment for Statistical Computing, 2015. URL `https://www.r-project.org/`.

S. Rabe-Hesketh and A. Skrondal. *Multilevel and Longitudinal Modeling Using Stata*. Stata Press, College Station, TX, 2nd edition, 2005.

S. Rasmussen and S. Abildstrom. Hospital variation in use of secondary preventive medicine after discharge for first acute myocardial infarction during 1995–2004. *Medical care*, 2008.

M. Redshaw and K. Hamilton. A survey of current neonatal unit organisation and policy. Technical report, National Perinatal Epidemiology Unit, Oxford, 2005. URL `https://npeu.ox.ac.uk/downloads/files/reports/Bliss-Final-Report.pdf`.

M. E. Redshaw and K. Hamilton. Networks, admissions and transfers: the perspectives of networks, neonatal units and parents. Technical report, National Perinatal Epidemiology Unit, Oxford, 2006.

M. J. Reeves, J. Gargano, K. S. Maier, J. P. Broderick, M. Frankel, K. A. LaBresh, C. J. Moomaw, and L. Schwamm. Patient-level and hospital-level determinants of the quality of acute stroke care: a multilevel modeling approach. *Stroke; a journal of cerebral circulation*, 41(12):2924–31, dec 2010. ISSN 1524-4628. doi: 10.1161/STROKEAHA.110.598664.

L. A. Reineck, F. Pike, T. Q. Le, B. D. Cicero, T. J. Iwashyna, and J. M. Kahn. Hospital factors associated with discharge bias in ICU performance measurement. *Critical care medicine*, 42(5):1055–64, may 2014. ISSN 1530-0293. doi: 10.1097/CCM.0000000000000132.

T. Reistetter, Y. Kuo, and A. Karmarkar. Geographic and facility variation in inpatient stroke rehabilitation: multilevel analysis of functional status. *Archives of physical*, 2015.

D. K. Richardson, J. D. Corcoran, G. J. Escobar, and S. K. Lee. SNAP-II and SNAPPE-II: Simplified newborn illness severity and mortality risk scores. *The Journal of Pediatrics*, 138(1):92–100, 2001. doi: 10.1067/mpd.2001.109608.

C. Roberts and R. Walwyn. Design and analysis of non-pharmacological treatment trials with multiple therapists per patient. *Statistics in Medicine*, page doi: 10.1002/sim.5521, 2012. doi: 10.1002/sim.5521.

K. Rothman, S. Greenland, and T. Lash. *Modern epidemiology*. 2008.

Royal College of Obstetricians and Gynaecologists. Antenatal Corticosteroids to Reduce Neonatal Morbidity and Mortality. Technical report, 2010. URL `https://www.rcog.org.uk/en/guidelines-research-services/guidelines/gtg7/`.

Royal College of Paediatrics and Child Health. National Neonatal Audit Programme 2013. Technical report, 2014. URL `http://www.rcpch.ac.uk/index.php?q=improving-child-health/quality-improvement-and-clinical-audit/national-neonatal-audit-programme-nnap`.

Royal College of Paediatrics and Child Health. National Neonatal Audit Programme Annual Report 2014, 2015. URL `http://www.rcpch.ac.uk/index.php?q=improving-child-health/quality-improvement-and-clinical-audit/national-neonatal-audit-programme-nnap`.

P. Royston and W. Sauerbrei. *Multivariable Model - Building: A Pragmatic Approach to Regression Anaylsis based on Fractional Polynomials for Modelling Continuous Variables*. John Wiley & Sons, 2008. ISBN 0470770783.

D. Rubin. Inference and missing data. *Biometrika*, 1976.

J. Rumsfeld, D. Magid, and E. Peterson. Outcomes after acute coronary syndrome admission to primary versus tertiary Veterans Affairs medical centers: The Veterans Affairs Access to Cardiology study. *American heart*, 2006.

M. A. Rysavy, L. Li, E. F. Bell, A. Das, S. R. Hintz, et al. Between-Hospital Variation in Treatment and Outcomes in Extremely Preterm Infants. *http://dx.doi.org/10.1056/NEJMoa1410689*, 2015.

S. Saigal and L. W. Doyle. An overview of mortality and sequelae of preterm birth from infancy to adulthood. *The Lancet*, 371(9608):261–269, 2008. doi: 10.1016/s0140-6736(08)60136-1.

Z. Samadashvili, E. Hannan, K. Cozzens, and G. Walford. Assessing Hospital Performance for Acute Myocardial Infarction: How Should Emergency Department Transfers Be Attributed. *Medical care*, 53(3):245–252, 2015.

O. Sauzet, K. C. Wright, L. Marston, P. Brocklehurst, and J. L. Peacock. Modelling the hierarchical structure in datasets with very small clusters: a simulation study to explore the effect of the proportion of clusters when the outcome is continuous. *Statistics in medicine*, 32(8):1429–38, apr 2013. ISSN 1097-0258. doi: 10.1002/sim.5638.

J. M. Schaaf, B. W. J. Mol, A. Abu-Hanna, and A. C. J. Ravelli. Trends in preterm birth: singleton and multiple pregnancies in the Netherlands, 2000-2007. *BJOG : an international journal of obstetrics and gynaecology*, 118(10):1196–204, sep 2011. ISSN 1471-0528. doi: 10.1111/j.1471-0528.2011.03010.x.

C. Seghieri, S. Mimmi, J. Lenzi, and M. P. Fantini. 30-day in-hospital mortality after acute myocardial infarction in Tuscany (Italy): an observational study using hospital discharge data. *BMC medical research methodology*, 12:170, 2012. ISSN 1471-2288. doi: 10.1186/1471-2288-12-170.

F. Serenius, G. Sjörs, M. Blennow, V. Fellman, G. Holmström, et al. EXPRESS study shows significant regional differences in 1-year outcome of extremely preterm infants in Sweden. *Acta paediatrica (Oslo, Norway : 1992)*, 103(1):27–37, jan 2014. ISSN 1651-2227. doi: 10.1111/apa.12421.

D. Sethi, S. Aljunid, S. Saperi, and F. Clemens. Comparison of the effectiveness of trauma services provided by secondary and tertiary hospitals in Malaysia. *Annals of emergency*, 49(1):52–61, 2007.

P. S. Shah, S. K. Lee, K. Lui, G. Sjörs, R. Mori, et al. The International Network for Evaluating Outcomes of very low birth weight, very preterm neonates (iNeo): a protocol for collaborative comparisons of international health services for quality improvement in neonatal care. *BMC Pediatrics*, 14(1):110, dec 2014. ISSN 1471-2431. doi: 10.1186/1471-2431-14-110.

D. M. Shahian and S.-L. T. Normand. What is a performance outlier? *BMJ quality & safety*, 24(2):95–9, mar 2015. ISSN 2044-5423. doi: 10.1136/bmjqs-2015-003934.

D. M. Shahian, D. F. Torchiana, R. J. Shemin, J. D. Rawn, and S.-L. T. Normand. Massachusetts Cardiac Surgery Report Card: Implications of Statistical Methodology. *The Annals of Thoracic Surgery*, 80(6):2106–2113, 2005. doi: 10.1016/j.athoracsur.2005.06.078.

C. Shapiro-Mendoza and E. Lackritz. Epidemiology of late and moderate preterm birth. *Seminars in Fetal and Neonatal Medicine*, 17(3):120–125, 2012.

W. Shen and T. Louis. Triple-goal estimates in two-stage hierarchical models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60(2):455–471, 1998.

P. A. Shlossman, J. S. Manley, A. C. Sciscione, and G. H. C. Colmorgen. An Analysis of Neonatal Morbidity and Mortality in Maternal (In Utero) and Neonatal Transports at 24-34 Weeks' Gestation. *Amer J Perinatol*, 14(08):449,456, 1997. doi: 10.1055/s-2007-994178.

P. Sidebotham, J. Fraser, P. Fleming, M. Ward-Platt, and R. Hain. Patterns of child death in England and Wales. *Lancet (London, England)*, 384(9946):904–14, sep 2014. ISSN 1474-547X. doi: 10.1016/S0140-6736(13)61090-9.

P. Silcocks and D. Kendrick. Spatial effects should be allowed for in primary care and other community-based cluster RCTS. *Trials*, 11:55, 2010.

M. M. Slattery and J. J. Morrison. Preterm delivery. *The Lancet*, 360(9344):1489–1497, 2002. ISSN 01406736. doi: 10.1016/S0140-6736(02)11476-0.

L. Smith, E. Draper, B. Manktelow, and D. Field. Socioeconomic inequalities in survival and provision of neonatal care: population based study of very preterm infants. *BMJ*, 339, 2009. doi: 10.1136/bmj.b4702.

T. A. B. Snijders and R. Bosker. *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. SAGE, London, 1999.

D. Spiegelhalter, L. Freedman, and M. Parmar. Bayesian approaches to randomized trials. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 157 (3):357–418, 1994.

D. Spiegelhalter, A. Thomas, N. Best, and D. Lunn. WinBUGS user manual. 2003.

D. Spiegelhalter, C. Sherlaw-Johnson, M. Bardsley, I. Blunt, C. Wood, and O. Grigg. Statistical methods for healthcare regulation: rating, screening and surveillance. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 175(1):1–47, jan 2012. ISSN 09641998. doi: 10.1111/j.1467-985X.2011.01010.x.

D. J. Spiegelhalter. Surgical Audit: Statistical Lessons from Nightingale and Codman. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 162(1):45–58, 1999.

D. J. Spiegelhalter. Handling over-dispersion of performance indicators. *Quality and Safety in Health Care*, 14(5):347–351, 2005a. doi: 10.1136/qshc.2005.013755.

D. J. Spiegelhalter. Funnel plots for comparing institutional performance. *Statistics in Medicine*, 24(8):1185–1202, 2005b.

D. J. Spiegelhalter, N. G. Best, B. P. Carlin, and A. Van Der Linde. Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(4):583–639, 2002. doi: 10.1111/1467-9868.00353.

StataCorp. Stata Statistical Software: Release 12, 2011.

Y. Statnikov and D. Gray. Neonatal Data Analysis Unit Data Management Document version 1.3. 2015.

E. W. Steyerberg, A. J. Vickers, N. R. Cook, T. Gerds, M. Gonen, N. Obuchowski, M. J. Pencina, and M. W. Kattan. Assessing the performance of prediction models: a framework for traditional and novel measures. *Epidemiology (Cambridge, Mass.)*, 21(1):128–38, jan 2010. ISSN 1531-5487. doi: 10.1097/EDE.0b013e3181c30fb2.

L. G. Suter, S.-X. Li, J. N. Grady, Z. Lin, Y. Wang, et al. National patterns of risk-standardized mortality and readmission after hospitalization for acute myocardial infarction, heart failure, and pneumonia: update on publicly reported outcomes measures based on the 2013 release. *Journal of general internal medicine*, 29(10):1333–40, oct 2014. ISSN 1525-1497. doi: 10.1007/s11606-014-2862-5.

C. Szmaragd, P. Clarke, and F. Steele. Subject specific and population average models for binary longitudinal data: a tutorial. *Longitudinal and Life Course Studies*, 4(2): 147–165, 2013. doi: 10.1136/bmj.b2393.

P. P. Tekkis, P. McCulloch, A. C. Steger, I. S. Benjamin, and J. D. Poloniecki. Mortality control charts for comparing performance of surgical units: validation study using hospital mortality data. *BMJ*, 326(7393):786, 2003. doi: 10.1136/bmj.326.7393.786.

The UK Neonatal Staffing Study Group. Patient volume, staffing, and workload in relation to risk-adjusted outcomes in a random stratified sample of UK neonatal intensive care units: a prospective evaluation. *The Lancet*, 359(9301):99–107, 2002.

N. Thomas, N. T. Longford, and J. E. Rolph. Empirical Bayes methods for estimating hospital-specific mortality rates. *Statistics in Medicine*, 13(9):889–903, 1994. doi: 10.1002/sim.4780130902.

C. V. Towers, R. Bonebrake, G. Padilla, and P. Rumney. The Effect of Transport on the Rate of Severe Intraventricular Hemorrhage in Very Low Birth Weight Infants. *Obstetrics & Gynecology*, 95(2):291–295, 2000.

J. Tucker and W. McGuire. Epidemiology of preterm birth. *BMJ*, 329(7467):675–678, 2004. doi: 10.1136/bmj.329.7467.675.

J. E. Tyson, N. A. Parikh, J. Langer, C. Green, and R. D. Higgins. Intensive Care for Extreme Prematurity — Moving beyond Gestational Age. *New England Journal of Medicine*, 358(16):1672–1681, 2008. doi: doi:10.1056/NEJMoa073059.

K. Walker, J. Neuburger, O. Groene, D. A. Cromwell, and J. van der Meulen. Public reporting of surgeon outcomes: low numbers of procedures lead to false complacency. *Lancet (London, England)*, 382(9905):1674–7, nov 2013. ISSN 1474-547X. doi: 10.1016/S0140-6736(13)61491-9.

S. I. Watson, W. Arulampalam, S. Petrou, N. Marlow, A. S. Morgan, E. S. Draper, S. Santhakumaran, and N. Modi. The effects of designation and volume of neonatal care on mortality and morbidity outcomes of very preterm infants in England: retrospective population-based cohort study. *BMJ open*, 4(7):e004856, jan 2014. ISSN 2044-6055. doi: 10.1136/bmjopen-2014-004856.

R. H. White, E. M. Geraghty, A. Brunson, S. Murin, T. Wun, F. Spencer, and P. S. Romano. High variation between hospitals in vena cava filter use for venous thromboembolism. *JAMA internal medicine*, 173(7):506–12, apr 2013. ISSN 2168-6114. doi: 10.1001/jamainternmed.2013.2352.

A. J. Wilcox. On the importance—and the unimportance— of birthweight. *International Journal of Epidemiology*, 30(6):1233–1241, 2001. doi: 10.1093/ije/30.6.1233.

A. R. Wilkinson, J. Ahluwalia, A. Cole, D. Crawford, J. Fyle, A. Gordon, J. Moorcraft, T. Pollard, and T. Roberts. Management of babies born extremely preterm at less than 26 weeks of gestation: a framework for clinical practice at the time of birth. *Archives of disease in childhood. Fetal and neonatal edition*, 94(1):F2–5, jan 2009. ISSN 1468-2052. doi: 10.1136/adc.2008.143321.

D. J. C. Wilkinson, P. Thiele, A. Watkins, and L. De Crespigny. Fatally flawed? A review and ethical analysis of lethal congenital malformations. *BJOG : an international journal of obstetrics and gynaecology*, 119(11):1302–8, oct 2012. ISSN 1471-0528. doi: 10.1111/j.1471-0528.2012.03450.x.

E. J. Wolf, A. M. Vintzileos, T. S. Rosenkrantz, J. F. Rodis, L. Lettieri, and A. Mallozzi. A comparison of pre-discharge survival and morbidity in singleton and twin very low birth weight infants. *Obstetrics and gynecology*, 80(3 Pt 1):436–9, sep 1992. ISSN 0029-7844.

L. Wolff-Smith and S. Beretvas. The impact of using incorrect weights with the multiple membership random effects model. *European Journal of Research Methods for the Behavioural and Social Sciences*, 10(1):31–42, 2014.

World Health Organization. Born Too Soon: The Global Action Report on Preterm Birth. Technical report, 2012. URL `http://www.who.int/pmnch/media/news/2012/preterm{\_}birth{\_}report/en/`.

C. M. Wright, A. F. Williams, D. Elliman, H. Bedford, E. Birks, G. Butler, M. Sachs, R. J. Moy, and T. J. Cole. Using the new UK-WHO growth charts. *BMJ*, 340, 2010. doi: 10.1136/bmj.c1140.

Y. Xu, Y. Liu, T. Shu, W. Yang, and M. Liang. Variations in the Quality of Care at Large Public Hospitals in Beijing, China: A Condition-Based Outcome Approach. *PloS one*, 10(10):e0138948, 2015. ISSN 1932-6203. doi: 10.1371/journal.pone.0138948.

X. Yang, B. Peng, R. Chen, Q. Zhang, D. Zhu, Q. J. Zhang, F. Xue, and L. Qi. Statistical profiling methods with hierarchical logistic regression for healthcare providers with binary outcomes. *Journal of Applied Statistics*, 41(1):46–59, aug 2013. ISSN 0266-4763. doi: 10.1080/02664763.2013.830086.

K. Yau and C. McGilchrist. ML and REML estimation in survival analysis with time dependent correlated frailty. *Statistics in medicine*, 17(11):1201–1213, 1998.

L. N. Yelland, T. R. Sullivan, and M. Makrides. Accounting for multiple births in randomised trials: a systematic review. *Archives of disease in childhood. Fetal and neonatal edition*, 100(2):F116–20, mar 2015. ISSN 1468-2052. doi: 10.1136/archdischild-2014-306239.

J. Zeitlin. Fetal sex and preterm birth: are males at greater risk? *Human Reproduction*, 17(10):2762–2768, oct 2002. ISSN 14602350. doi: 10.1093/humrep/17.10.2762.

J. Zeitlin, K. Szamotulska, N. Drewniak, A. Mohangoo, J. Chalmers, L. Sakkeus, L. Irgens, M. Gatt, M. Gissler, and B. Blondel. Preterm birth time trends in Europe: a study of 19 countries. *BJOG: An International Journal of Obstetrics & Gynaecology*, 120(11):1356–1365, oct 2013. ISSN 14700328. doi: 10.1111/1471-0528.12281.

Appendix A

List of neonatal units included in the thesis

The table below lists all the neonatal units (NNUs) in England with at least one admission of an infant born below 32 weeks gestation between 2011 and 2013 included in the National Neonatal Research Database (NNRD) and therefore included in this thesis. Note that some NNUs did not contribute data for the whole period due to closures and mergers. NNUs are grouped by neonatal network, and the level of NNU (special care baby unit (SCBU), local neonatal unit (LNU) and neonatal intensive care unit (NICU), see section 1.2 for a description of these classifications).

Table A.1: Neonatal units contributing data to the National Neonatal Research Database and included in the thesis *(continues to table A.2)*

Neonatal network	Neonatal Unit	Level
Bedfordshire and Hertfordshire	Bedford Hospital	SCBU
	Queen Elizabeth II Hospital	SCBU
	Watford General Hospital	LNU
	Lister Hospital	LNU
	Luton & Dunstable Hospital	NICU
Cheshire and Merseyside	Whiston Hospital	LNU
	Leighton Hospital	LNU
	Macclesfield District General Hospital	LNU
	Countess Of Chester Hospital	LNU
	Ormskirk District General Hospital	LNU
	Warrington Hospital	LNU
	Arrowe Park Hospital	NICU
	Liverpool Women's Hospital	NICU
Eastern	James Paget Hospital	SCBU
	West Suffolk Hospital	SCBU
	Hinchingbrooke Hospital	SCBU
	Queen Elizabeth Hospital, King's Lynn	LNU
	Colchester General Hospital	LNU
	Peterborough City Hospital	LNU
	Ipswich Hospital	LNU
	Princess Alexandra Hospital	LNU
	Broomfield Hospital	LNU
	Rosie Maternity Hospital, Addenbrookes	NICU
	Norfolk & Norwich University Hospital	NICU
Greater Manchester	University Hospital Of South Manchester	LNU
	Tameside General Hospital	LNU
	Royal Albert Edward Infirmary	LNU
	Fairfield General Hospital	LNU
	Rochdale Infirmary	LNU
	North Manchester General Hospital	LNU
	Stepping Hill Hospital	LNU
	Salford Royal Hospital	NICU
	Royal Bolton Hospital	NICU
	Royal Oldham Hospital	NICU
	St Mary's Hospital, Manchester	NICU
Kent	Darent Valley Hospital	SCBU
	Maidstone General Hospital	SCBU
	Queen Elizabeth The Queen Mother Hospital	SCBU
	Tunbridge Wells Hospital	LNU
	Medway Maritime Hospital	NICU
	William Harvey Hospital	NICU
Lancashire and South Cumbria	Furness General Hospital	SCBU
	Royal Lancaster Infirmary	LNU
	Victoria Hospital, Blackpool	LNU
	Royal Preston Hospital	NICU
	Lancashire Women & Newborn Centre	NICU

Table A.2: Neonatal units contributing data to the National Neonatal Research Database and included in the thesis (*continued from table A.1*)

Neonatal network	Neonatal Unit	Level
London - North Central	The Royal Free Hospital	SCBU
	Chase Farm Hospital	SCBU
	Barnet Hospital	LNU
	Whittington Hospital	LNU
	University College Hospital	NICU
London - North East	King George Hospital	SCBU
	Southend Hospital	LNU
	Basildon Hospital	LNU
	North Middlesex University Hospital	LNU
	Queen's Hospital, Romford	LNU
	Whipps Cross University Hospital	LNU
	Newham General Hospital	LNU
	The Royal London Hospital - Constance Green	NICU
	The Royal London Hospital	NICU
	Homerton Hospital	NICU
London - North West	Ealing Hospital	SCBU
	West Middlesex University Hospital	SCBU
	Hillingdon Hospital	LNU
	Northwick Park Hospital	LNU
	St Mary's Hospital, London	LNU
	Chelsea & Westminster Hospital	NICU
	Queen Charlotte's Hospital	NICU
London - South East	Princess Royal University Hospital	SCBU
	Queen Elizabeth Hospital, Woolwich	LNU
	University Hospital Lewisham	LNU
	Guy's & St Thomas' Hospital	NICU
	King's College Hospital	NICU
London - South West	Epsom General Hospital	SCBU
	Croydon University Hospital	LNU
	Kingston Hospital	LNU
	St Helier Hospital	LNU
	St George's Hospital	NICU
Midlands - Central	Warwick Hospital	SCBU
	Queen's Hospital, Burton On Trent	SCBU
	George Eliot Hospital	SCBU
	Kettering General Hospital	LNU
	Northampton General Hospital	LNU
	University Hospital Coventry	NICU
	Leicester Neonatal Service	NICU

Table A.3: Neonatal units contributing data to the National Neonatal Research Database and included in the thesis (*continued from table A.2*)

Neonatal network	Neonatal Unit	Level
Midlands - South West	Good Hope Hospital	SCBU
	Alexandra Hospital	SCBU
	Hereford County Hospital	SCBU
	Worcestershire Royal Hospital	LNU
	City Hospital, Birmingham	LNU
	Birmingham Women's Hospital	NICU
	Birmingham Heartlands Hospital	NICU
Midlands North SSBC*	County Hospital, Staffordshire	SCBU
	Manor Hospital	LNU
	Russells Hall Hospital	LNU
	Royal Shrewsbury Hospital	LNU
	University Hospital Of North Staffordshire	NICU
	New Cross Hospital	NICU
North Trent	Bassetlaw District General Hospital	SCBU
	Barnsley District General Hospital	LNU
	Rotherham District General Hospital	LNU
	Chesterfield & North Derbyshire Royal Hospital	LNU
	Doncaster Royal Infirmary	LNU
	Diana Princess Of Wales Hospital	LNU
	Scunthorpe General Hospital	LNU
	The Jessop Wing, Sheffield	NICU
Northern	South Tyneside District Hospital	SCBU
	Queen Elizabeth Hospital, Gateshead	SCBU
	Wansbeck General Hospital	SCBU
	Darlington Memorial Hospital	SCBU
	Cumberland Infirmary	SCBU
	West Cumberland Hospital	SCBU
	University Hospital Of North Durham	SCBU
	Sunderland Royal Hospital	NICU
	Royal Victoria Infirmary	NICU
	James Cook University Hospital	NICU
	University Hospital Of North Tees	NICU
Peninsula - South West	Torbay Hospital	SCBU
	North Devon District Hospital	SCBU
	Royal Devon & Exeter Hospital	LNU
	Royal Cornwall Hospital	LNU
	Derriford Hospital	NICU

* Midlands North, Staffordshire, Shropshire and Black Country

Table A.4: Neonatal units contributing data to the National Neonatal Research Database and included in the thesis (*continued from table A.3*)

Neonatal network	Neonatal Unit	Level
South Central (North)	Horton Hospital	SCBU
	Wexham Park Hospital	LNU
	Milton Keynes Hospital	LNU
	Royal Berkshire Hospital	LNU
	Stoke Mandeville Hospital	LNU
	John Radcliffe Hospital	NICU
South Central (South)	Dorset County Hospital	LNU
	Poole Hospital	LNU
	Royal Hampshire County Hospital	LNU
	Basingstoke & North Hampshire Hospital	LNU
	Salisbury District Hospital	LNU
	St Richard's Hospital	LNU
	St Mary's Hospital, Isle of Wight	LNU
	Princess Anne Hospital	NICU
	Queen Alexandra Hospital	NICU
Surrey and Sussex	Royal Surrey County Hospital	SCBU
	Frimley Park Hospital	SCBU
	Worthing Hospital	SCBU
	Princess Royal Hospital	SCBU
	Conquest Hospital	SCBU
	Eastbourne District General Hospital	SCBU
	East Surrey Hospital	LNU
	St Peter's Hospital	NICU
	Royal Sussex County Hospital	NICU
Trent	Pilgrim Hospital	SCBU
	Lincoln County Hospital	LNU
	King's Mill Hospital	LNU
	Royal Derby Hospital	LNU
	Nottingham City Hospital	NICU
	Nottingham University Hospital	NICU

Table A.5: Neonatal units contributing data to the National Neonatal Research Database and included in the thesis (*continued from table A.4*)

Neonatal network	Neonatal Unit	Level
Western	Yeovil District Hospital	SCBU
	Taunton & Somerset Hospital	LNU
	Royal United Hospital	LNU
	Great Western Hospital	LNU
	Gloucestershire Royal Hospital	LNU
	St Michael's Hospital	NICU
	Southmead Hospital	NICU
Yorkshire	Scarborough General Hospital	SCBU
	Harrogate District Hospital	SCBU
	Dewsbury & District Hospital	SCBU
	York District Hospital	LNU
	Airedale General Hospital	LNU
	Pinderfields General Hospital	LNU
	Calderdale Royal Hospital	LNU
	Bradford Royal Infirmary	NICU
	Leeds Neonatal Service	NICU
	Hull Royal Infirmary	NICU

Appendix B

Updating the search to identify risk factors for mortality of preterm infants

Medlock et al. (2011) carried out a systematic review of prediction models for mortality of preterm infants, one of the aims of which was to identify important predictor variables. As described in section 3.3.1, I selected infant risk factors to adjust for in the thesis based on the findings of the review. The authors searched the MedLine database for studies which provided mortality predictions for infants less than 32 weeks gestation and/or less than 1500g birth weight, or subpopulations of this group provided they did not have a specific disease or condition. Full methods for the study and the resulting papers can be found in the published paper. I updated the search using the same search strategy as there may have been further developments since publication of the study. The resulting studies, including a brief description and the variables included are shown in tables B.1 and B.2. All variables included in prediction models of the studies in tables B.1 and B.2 were either already considered by Medlock et al., or only appeared in one or two studies so were not considered further.

Table B.1: Results from updating the search of Medlock et al. (2011), continues to table B.2

First author	Year published	Birth Year	Country	Cohort	Excluded congenital malformations	Deaths/total (%)	Outcome	Variables included
Gooden	2014	2005-2006	Jamaica	Admissions GA<32w or BWT<1500g to one NNU	Yes	58/109 (53)	Death before discharge	CRIB II, GA, BWT, sex
Mukhopadhyay	2013	2009-2011	India	Admissions GA>23w and BWT 400-999g to one NNU	Yes	68/149 (46)	Death before discharge	BWT, mechanical ventilation, hypotensive shock
Fatttore	2015	2006-2008	EuroHOPE	All livebirths in population BWT<1500g or GA<32w	Yes	4.7% (Scotland) to 15.7% (Hungary)	Death within 30 days	GA, Apgar5, sex, multiple birth, congenital malformations, income, pop^n density,
Manktelow	2013	2008-2010	UK	White singleton admissions GA 23-32 weeks in two regions	Yes	244/2995 (8)	Death before discharge	GA, BWT, sex
Ambalavanan	2012	1998-2005	USA	Admissions GA 22-32w and BWT 401-1000g to 18 NNUs	Yes	4448/13085 (34%)	Death before discharge	BWT, Apgar5, sex, GA, steroids
Kugelman	2012	1995-2008	Israel	Livebirths GA 23-26w from all hospitals	No	2441/4408 (55)	Death before discharge	GA, BWT, steroids, sex
Cetinkaya	2014	2008-2012	Turkey	Livebirths BWT<1500g at one hospital	Yes	56/241 (23)	Death before discharge	GA, BWT, steroids, Doppler findings
Dong	2012	2002-2009	China	Admissions GA<37w and BWt<1500g to one NNU	Yes	136/475 (29)	Death before discharge	Persistent pulmonary hypertension of newborn, pulmonary haemorrhage, BWT, GA, feeding before 3 postnatal days and enteral feeding
Shah	2012	2003-2008	Canada	Admissions GA 22-32w to all NNUs	Yes	6028/17148 (35)	Death before discharge	Sex, GA, BWT, steroids, multiple birth, mode of delivery, smoking.
Salas	2016	1998-2009	USA	Livebirths GA<26w at 18 hospitals	Yes	Not stated	Death before 18-22 month follow up visit	Birth year, sex, ethnicity, steroids, multiple birth, pre-eclampsia, Apgar5, SGA
Vincer	2015	1996-2012	Canada	All livebirths GA 23-31w in one region	Yes	124/1074 (12)	Death within 28 days	GA, SGA, oligohydramnios, maternal psychiatric disorders, antenatal antibiotics, monochorionic twins, antihypertensive agents, steroids
Ravelli	2014	2000-2007	The Netherlands	All livebirths GA 25-31 and BWT >500g in the country	Yes	766/8500 (9)	Death within 28 days	GA, steroids, sex , maternal age, ethnicity, presentation at delivery and level of hospital
Pepler	2012	2007-2008	Southern Africa	Livebirths GA 22-29w or BWT 400-1500g from 15 hospitals	No	91/2307 (4)	Death before discharge	Apgar1, BW, mode of delivery

Abbreviations: Apgar1(5) Apgar score at 1 (5) minutes; BWT Birth weight; CRIB Clinical Index of Risk for Babies; GA Gestational age; PROM Premature Rupture of Membranes; SGA Small for gestational age; SNAPPE Score for Neonatal Acute Physiology Perinatal Extension
EuroHOPE countries/regions are Finland, Hungary, province of Rome, the Netherlands, Norway, Scotland and Sweden

Table B.2: Results from updating the search of Medlock et al. (2011), continued from table B.1

First author	Year published	Birth Year	Country	Cohort	Excluded congenital malformations	Deaths/total (%)	Outcome	Variables included
Ge	2013	2010-2011	Canada	Admissions GA 23-30w to all NNUs	Yes	611/6424 (10)	Death before discharge	GA, SGA, sex, SNAPPE II, outborn status, steroids, surfactant, mechanical ventilation
Ogawa	2013	2001-2002	Japan	Livebirths BWT<1000g and GA 22-36w from 133 hospitals	Yes	366/1713 (21)	Death before discharge	Placental abruption, steroids, presentation at delivery, mode of delivery, birth weight, Apgar5.
Reid	2014	2003-2004	Australia	Admissions GA<32w in two regions	Yes	8.5%	Death before discharge	Compared CRIB and SNAPPE scores, performed similarly and no other variables were predictive
Schenone	2010	2000-2007	USA	Singleton livebirths GA 22-25w at one hospitals	Yes	117/179 (65)	Death before discharge and within 120 days	Femur length, GA, estimated fetal weight, PROM, steroids
Wang	2011	2004-2007	Norway	Livebirths GA 24-27w at one hospital	No	21/109 (19)	Death before 28 days/discharge	Multiple pregnancy, GA
Lee	2010	2005-2008	USA	Admissions GA 22-25w and BWT 401-1000g to 128 NNUs	Yes	47%	Death before discharge	GA, BW, steroids, sex, multiple birth

Abbreviations: Apgar1(5) Apgar score at 1 (5) minutes; BWT Birth weight; CRIB Clinical Index of Risk for Babies; GA Gestational age; PROM Premature Rupture of Membranes; SGA Small for gestational age; SNAPPE Score for Neonatal Acute Physiology Perinatal Extension

References

N. Ambalavanan, W. A. Carlo, J. E. Tyson, J. C. Langer, M. C. Walsh, et al. Outcome trajectories in extremely preterm infants. *Pediatrics*, 130(1):e115–25, jul 2012. ISSN 1098-4275. doi: 10.1542/peds.2011-3693.

S. E. Cetinkaya, E. Okulu, F. Soylemez, I. M. Akin, S. Sahin, T. Akyel, S. Alan, B. Atasay, S. Arsan, and A. Koc. Perinatal risk factors and mode of delivery associated with mortality in very low birth weight infants. *The journal of maternal-fetal & neonatal medicine : the official journal of the European Association of Perinatal Medicine, the Federation of Asia and Oceania Perinatal Societies, the International Society of Perinatal Obstetricians*, pages 1–6, sep 2014. ISSN 1476-4954. doi: 10.3109/14767058.2014.953476.

Y. Dong, G. Yue, and J.-L. Yu. Changes in perinatal care and predictors of in-hospital mortality for very low birth weight preterm infants. *Iranian journal of pediatrics*, 22 (3):326–32, sep 2012. ISSN 2008-2142.

G. Fatttore, D. Numerato, M. Peltola, H. Banks, R. Graziani, et al. Variations and Determinants of Mortality and Length of Stay of Very Low Birth Weight and Very Low for Gestational Age Infants in Seven European Countries. *Health economics*, 24 Suppl 2:65–87, dec 2015. ISSN 1099-1050. doi: 10.1002/hec.3261.

W. J. Ge, L. Mirea, J. Yang, K. L. Bassil, S. K. Lee, P. S. Shah, and Canadian Neonatal Network. Prediction of neonatal outcomes in extremely preterm neonates. *Pediatrics*, 132(4):e876–85, oct 2013. ISSN 1098-4275. doi: 10.1542/peds.2013-0702.

M. Gooden, N. Younger, and H. Trotman. What is the best predictor of mortality in a very low birth weight infant population with a high mortality rate in a medical setting with limited resources? *American journal of perinatology*, 31(6):441–6, jun 2014. ISSN 1098-8785. doi: 10.1055/s-0033-1351658.

A. Kugelman, D. Bader, L. Lerner-Geva, V. Boyko, O. Levitzki, A. Riskin, and B. Reichman. Poor outcomes at discharge among extremely premature infants: a national population-based study. *Archives of pediatrics & adolescent medicine*, 166(6):543–50, jun 2012. ISSN 1538-3628. doi: 10.1001/archpediatrics.2011.891.

H. C. Lee, C. Green, S. R. Hintz, J. E. Tyson, N. A. Parikh, J. Langer, and J. B. Gould. Prediction of Death for Extremely Premature Infants in a Population-Based Cohort. *Pediatrics*, 126(3):e644–650, 2010. doi: 10.1542/peds.2010-0097.

B. N. Manktelow, S. Seaton, D. Field, and E. Draper. Population based estimates of in-unit survival for very preterm babies. *Pediatrics*, 131(2):e425–32, 2013.

K. Mukhopadhyay, D. Louis, R. Mahajan, and P. Kumar. Predictors of mortality and major morbidities in extremely low birth weight neonates. *Indian pediatrics*, 50(12): 1119–23, dec 2013. ISSN 0974-7559.

M. Ogawa, Y. Matsuda, E. Kanda, J. Konno, M. Mitani, Y. Makino, and H. Matsui. Survival rate of extremely low birth weight infants and its risk factors: case-control study in Japan. *ISRN obstetrics and gynecology*, 2013:873563, 2013. ISSN 2090-4436. doi: 10.1155/2013/873563.

P. T. Pepler, D. W. Uys, and D. G. Nel. Predicting mortality and length-of-stay for neonatal admissions to private hospital neonatal intensive care units: a Southern African retrospective study. *African health sciences*, 12(2):166–73, jun 2012. ISSN 1729-0503. doi: 10.4314/ahs.v12i2.14.

A. C. Ravelli, J. M. Schaaf, B. W. J. Mol, P. Tamminga, M. Eskes, J. A. van der Post, and A. Abu-Hanna. Antenatal prediction of neonatal mortality in very premature infants. *European Journal of Obstetrics & Gynecology and Reproductive Biology*, 176: 126–131, 2014. ISSN 03012115. doi: 10.1016/j.ejogrb.2014.02.030.

S. Reid, B. Bajuk, K. Lui, and E. A. Sullivan. Comparing CRIB-II and SNAPPE-II as mortality predictors for very preterm infants. *Journal of Paediatrics and Child Health*, 51(5):524–528, may 2015. ISSN 10344810. doi: 10.1111/jpc.12742.

A. A. Salas, W. A. Carlo, N. Ambalavanan, T. L. Nolen, B. J. Stoll, A. Das, R. D. Higgins, and Eunice Kennedy Shriver National Institute of Child Health and Human Development Neonatal Research Network. Gestational age and birthweight for risk assessment of neurodevelopmental impairment or death in extremely preterm infants. *Archives of disease in childhood. Fetal and neonatal edition*, feb 2016. ISSN 1468-2052. doi: 10.1136/archdischild-2015-309670.

M. H. Schenone, E. Aguin, Y. Li, C. Lee, M. Kruger, and R. O. Bahado-Singh. Prenatal prediction of neonatal survival at the borderline viability. *The journal of maternal-fetal & neonatal medicine : the official journal of the European Association of Perinatal Medicine, the Federation of Asia and Oceania Perinatal Societies, the International Society of Perinatal Obstetricians*, 23(12):1413–8, dec 2010. ISSN 1476-4954. doi: 10.3109/14767058.2010.481318.

P. S. Shah, X. Y. Ye, A. Synnes, N. Rouvinez-Bouali, W. Yee, and S. K. Lee. Prediction of survival without morbidity for infants born at under 33 weeks gestational age: a user-friendly graphical tool. *Archives of Disease in Childhood - Fetal and Neonatal Edition*, 97(2):F110–F115, 2012. doi: 10.1136/archdischild-2011-300143.

M. J. Vincer, B. A. Armson, V. M. Allen, A. C. Allen, D. A. Stinson, et al. An Algorithm for Predicting Neonatal Mortality in Threatened Very Preterm Birth. *Journal of*

Obstetrics and Gynaecology Canada, 37(11):958–965, nov 2015. ISSN 17012163. doi: 10.1016/S1701-2163(16)30045-7.

Y. Wang, T. Tanbo, L. Ellingsen, T. Abyholm, and T. Henriksen. Effect of pregestational maternal, obstetric and perinatal factors on neonatal outcome in extreme prematurity. *Archives of gynecology and obstetrics*, 284(6):1381–7, dec 2011. ISSN 1432-0711. doi: 10.1007/s00404-011-1870-5.

Appendix C

Search results from literature review of transfers

W. Aelvoet, N. Terryn, A. Blommaert, G. Molenberghs, N. Hens, F. De Smet, M. Callens, and P. Beutels. Community-acquired pneumonia (CAP) hospitalizations and deaths: is there a role for quality improvement through inter-hospital comparisons? *International journal for quality in health care : journal of the International Society for Quality in Health Care / ISQua*, 28(1):22–32, feb 2016. ISSN 1464-3677. doi: 10.1093/intqhc/mzv092.

G. Ben-Josef, L. S. Ott, S. B. Spivack, C. Wang, J. S. Ross, S. J. Shah, J. P. Curtis, N. Kim, H. M. Krumholz, and S. M. Bernheim. Payments for acute myocardial infarction episodes-of-care initiated at hospitals with and without interventional capabilities. *Circulation. Cardiovascular quality and outcomes*, 7(6):882–8, nov 2014. ISSN 1941-7705. doi: 10.1161/CIRCOUTCOMES.114.000927.

J. G. Berry, S. L. Toomey, A. M. Zaslavsky, A. K. Jha, M. M. Nakamura, et al. Pediatric readmission prevalence and variability across hospitals. *JAMA*, 309(4):372–80, jan 2013. ISSN 1538-3598. doi: 10.1001/jama.2012.188351.

A. Bottle, R. Sanders, A. Mozid, and P. Aylin. Provider profiling models for acute coronary syndrome mortality using administrative data. *International journal of cardiology*, 168(1):338–343, 2013.

R. Bradbury, F. S. Jr, and P. Steen. Interhospital variations in admission severity-adjusted hospital mortality and morbidity. *Health services research*, 1991.

D. W. Bratzler, S.-L. T. Normand, Y. Wang, W. J. O'Donnell, M. Metersky, L. F. Han, M. T. Rapp, and H. M. Krumholz. An administrative claims model for profiling hospital 30-day mortality rates for pneumonia patients. *PloS one*, 6(4):e17401, 2011. ISSN 1932-6203. doi: 10.1371/journal.pone.0017401.

J. Bruckel, X. Liu, S. F. Hohmann, A. S. Karson, E. Mort, and D. M. Shahian. The denominator problem: national hospital quality measures for acute myocar-

dial infarction. *BMJ quality & safety*, mar 2016. ISSN 2044-5423. doi: 10.1136/bmjqs-2015-004888.

M. J. Campbell, R. M. Jacques, J. Fotheringham, R. Maheswaran, and J. Nicholl. Developing a summary hospital mortality index: retrospective analysis in English hospitals over five years. *BMJ (Clinical research ed.)*, 344:e1001, 2012. ISSN 1756-1833.

T. Chen, M. Ma, F. Chen, F. Hu, and Y. Lu. The relationship between survival after out-of-hospital cardiac arrest and process measures for emergency medical service ambulance team performance. *Resuscitation*, 2015.

I. Dégano, I. Subirana, M. Torre, M. Grau, and J. Vila. A European benchmarking system to evaluate in-hospital mortality rates in acute coronary syndrome: The EURHOBOP project. *International journal of*, 182:509–516, 2015.

D. Gomez, A. Alali, W. Xiong, B. Zarzaur, and N. Mann. Definitive care in level 3 trauma centres after severe injury: A comparison of clinical characteristics and outcomes. *Injury*, 46(9):1790–1795, 2015.

J. S. Goodwin, Y.-L. Lin, S. Singh, and Y.-F. Kuo. Variation in length of stay and outcomes among hospitalized patients attributable to hospitals and hospitalists. *Journal of general internal medicine*, 28(3):370–6, mar 2013. ISSN 1525-1497. doi: 10.1007/s11606-012-2255-6.

S. Hassani, A. Lindman, D. Kristoffersen, and O. Tomic. 30-Day Survival Probabilities as a Quality Indicator for Norwegian Hospitals: Data Management and Analysis. *PloS one*, 2015. doi: 10.1371/journal.pone.0136547.

A. F. Hernandez, G. C. Fonarow, L. Liang, P. A. Heidenreich, C. Yancy, and E. D. Peterson. The need for multiple measures of hospital quality: results from the Get with the Guidelines-Heart Failure Registry of the American Heart Association. *Circulation*, 124(6):712–9, aug 2011. ISSN 1524-4539. doi: 10.1161/CIRCULATIONAHA.111.026088.

R. Y. Hsia, S. M. Asch, R. E. Weiss, D. Zingmond, G. Gabayan, L.-J. Liang, W. Han, H. McCreath, and B. C. Sun. Is emergency department crowding associated with increased "bounceback" admissions? *Medical care*, 51(11):1008–14, nov 2013. ISSN 1537-1948. doi: 10.1097/MLR.0b013e3182a98310.

S. Jan, G. Slap, K. Smith-Whitley, D. Dai, R. Keren, and D. M. Rubin. Association of hospital and provider types on sickle cell disease outcomes. *Pediatrics*, 132(5):854–61, nov 2013. ISSN 1098-4275. doi: 10.1542/peds.2013-0089.

J. Jollis and P. Romano. Pennsylvania's focus on heart attack—grading the scorecard. *New England Journal of Medicine*, 338:983–987, 1998.

P. S. Keenan, S.-L. T. Normand, Z. Lin, E. E. Drye, K. R. Bhat, et al. An administrative claims measure suitable for profiling hospital performance on the basis of 30-day all-cause readmission rates among patients with heart failure. *Circulation. Cardiovascular quality and outcomes*, 1(1):29–37, sep 2008. ISSN 1941-7705. doi: 10.1161/CIRCOUTCOMES.108.802686.

N. Kim, S. M. Bernheim, L. S. Ott, L. Han, S. B. Spivack, X. Xu, M. Volpe, A. Liu, and H. M. Krumholz. An administrative claims measure of payments made for Medicare patients for a 30-day episode of care for acute myocardial infarction. *Medical care*, 53 (6):542–9, jun 2015. ISSN 1537-1948. doi: 10.1097/MLR.0000000000000361.

M. Kosseim, N. E. Mayo, S. Scott, J. A. Hanley, J. Brophy, B. Gagnon, and L. Pilote. Ranking hospitals according to acute myocardial infarction mortality: should transfers be included? *Medical Care*, 44(7):664–670, 2006.

H. M. Krumholz, R. G. Brindis, J. E. Brush, D. J. Cohen, A. J. Epstein, et al. Standards for Statistical Models Used for Public Reporting of Health Outcomes. *Circulation*, 113(3):456–462, 2006. doi: 10.1161/circulationaha.105.170769.

D. J. LaPar, I. L. Kron, D. R. Jones, G. J. Stukenborg, and B. D. Kozower. Hospital procedure volume should not be used as a measure of surgical quality. *Annals of surgery*, 256(4):606–15, oct 2012. ISSN 1528-1140. doi: 10.1097/SLA.0b013e31826b4be6.

A. H. Leyland and F. A. Boddy. League tables and acute myocardial infarction. *The Lancet*, 351(9102):555–558, 1998. doi: 10.1016/s0140-6736(97)09362-8.

J. H. Lichtman, E. C. Leifheit-Limson, S. B. Jones, Y. Wang, and L. B. Goldstein. Preventable readmissions within 30 days of ischemic stroke among Medicare beneficiaries. *Stroke; a journal of cerebral circulation*, 44(12):3429–35, dec 2013. ISSN 1524-4628. doi: 10.1161/STROKEAHA.113.003165.

J. P. Marcin and P. S. Romano. Impact of between-hospital volume and within-hospital volume on mortality and readmission rates for trauma patients in California. *Critical care medicine*, 32(7):1477–83, jul 2004. ISSN 0090-3493.

L. Moore, J. A. Hanley, A. F. Turgeon, A. Lavoie, and B. Eric. A new method for evaluating trauma centre outcome performance: TRAM-adjusted mortality estimates. *Annals of surgery*, 251(5):952–8, may 2010. ISSN 1528-1140. doi: 10.1097/SLA.0b013e3181d97589.

T. Osler, L. G. Glance, and D. W. Hosmer. Comparison of hospital mortality rates after burn injury in New York State: a risk-adjusted population-based observational study. *The Journal of trauma*, 71(4):1040–7, oct 2011. ISSN 1529-8809. doi: 10.1097/TA.0b013e318217866f.

J. Profit, L. A. Petersen, M. C. McCormick, G. J. Escobar, K. Coleman-Phox, Z. Zheng, K. Pietz, and J. A. F. Zupancic. Patient-to-nurse ratios and outcomes of moderately preterm infants. *Pediatrics*, 125(2):320–6, feb 2010. ISSN 1098-4275. doi: 10.1542/peds.2008-3140.

S. Rasmussen and S. Abildstrom. Hospital variation in use of secondary preventive medicine after discharge for first acute myocardial infarction during 1995–2004. *Medical care*, 2008.

M. J. Reeves, J. Gargano, K. S. Maier, J. P. Broderick, M. Frankel, K. A. LaBresh, C. J. Moomaw, and L. Schwamm. Patient-level and hospital-level determinants of the quality of acute stroke care: a multilevel modeling approach. *Stroke; a journal of cerebral circulation*, 41(12):2924–31, dec 2010. ISSN 1524-4628. doi: 10.1161/STROKEAHA.110.598664.

L. A. Reineck, F. Pike, T. Q. Le, B. D. Cicero, T. J. Iwashyna, and J. M. Kahn. Hospital factors associated with discharge bias in ICU performance measurement. *Critical care medicine*, 42(5):1055–64, may 2014. ISSN 1530-0293. doi: 10.1097/CCM.0000000000000132.

T. Reistetter, Y. Kuo, and A. Karmarkar. Geographic and facility variation in inpatient stroke rehabilitation: multilevel analysis of functional status. *Archives of physical*, 2015.

J. Rumsfeld, D. Magid, and E. Peterson. Outcomes after acute coronary syndrome admission to primary versus tertiary Veterans Affairs medical centers: The Veterans Affairs Access to Cardiology study. *American heart*, 2006.

Z. Samadashvili, E. Hannan, K. Cozzens, and G. Walford. Assessing Hospital Performance for Acute Myocardial Infarction: How Should Emergency Department Transfers Be Attributed. *Medical care*, 53(3):245–252, 2015.

C. Seghieri, S. Mimmi, J. Lenzi, and M. P. Fantini. 30-day in-hospital mortality after acute myocardial infarction in Tuscany (Italy): an observational study using hospital discharge data. *BMC medical research methodology*, 12:170, 2012. ISSN 1471-2288. doi: 10.1186/1471-2288-12-170.

D. Sethi, S. Aljunid, S. Saperi, and F. Clemens. Comparison of the effectiveness of trauma services provided by secondary and tertiary hospitals in Malaysia. *Annals of emergency*, 49(1):52–61, 2007.

B. Shah, A. Hernandez, and L. Liang. Hospital variation and characteristics of implantable cardioverter-defibrillator use in patients with heart failure: data from the GWTG-HF (Get With The. *Journal of the*, 2009.

L. Slobbe and O. Arah. Mortality in Dutch hospitals: trends in time, place and cause of death after admission for myocardial infarction and stroke. An observational study. *BMC health*, 2008.

L. G. Suter, S.-X. Li, J. N. Grady, Z. Lin, Y. Wang, et al. National patterns of risk-standardized mortality and readmission after hospitalization for acute myocardial infarction, heart failure, and pneumonia: update on publicly reported outcomes measures based on the 2013 release. *Journal of general internal medicine*, 29(10):1333–40, oct 2014. ISSN 1525-1497. doi: 10.1007/s11606-014-2862-5.

W. Weller, G. Kabra, and K. Cozzens. Appending limited clinical data to an administrative database for assessing institutional quality outliers for AMI care: Does it matter? *International journal of*, 2010.

R. H. White, E. M. Geraghty, A. Brunson, S. Murin, T. Wun, F. Spencer, and P. S. Romano. High variation between hospitals in vena cava filter use for venous thromboembolism. *JAMA internal medicine*, 173(7):506–12, apr 2013. ISSN 2168-6114. doi: 10.1001/jamainternmed.2013.2352.

Y. Xu, Y. Liu, T. Shu, W. Yang, and M. Liang. Variations in the Quality of Care at Large Public Hospitals in Beijing, China: A Condition-Based Outcome Approach. *PloS one*, 10(10):e0138948, 2015. ISSN 1932-6203. doi: 10.1371/journal.pone.0138948.

J. Yazdany, B. J. Marafino, M. L. Dean, N. S. Bardach, R. Duseja, M. M. Ward, and R. A. Dudley. Thirty-day hospital readmissions in systemic lupus erythematosus: predictors and hospital- and state-level variation. *Arthritis & rheumatology (Hoboken, N.J.)*, 66(10):2828–36, oct 2014. ISSN 2326-5205. doi: 10.1002/art.38768.

Appendix D

OpenBUGS model code

This appendix contains the OpenBUGSOpenBUGS Project Management Group (2014) code to specify the six single membership Bayesian hierarchical models and the multiple membership model described in chapter 5. Note that in OpenBUGS distributions are specified in terms of the precision, which is the inverse of the variance. In addition all regression coefficients were specified in the code in a single vector `beta` for convenience, rahter than using than using separate Greek letters for the intercept, infant, and NNU characteristics as described in chapter 5. Calculation of the weights for the multiple membership models was performed outside of OpenBUGS and separate data files created for each weighting with the same names, so the same code could be used to run the analysis for all weights.

```
############################################################
################          Model 1        ##################
############################################################

model{
    for (i in 1:N) {
        died[i] ~ dbern(p[i])
        logit(p[i]) <- beta[2]*ga1[i] + beta[3]*ga2[i] +
        beta[4]*bwt1[i] +  beta[5]*bwt2[i] +
        beta[6]*ga1bwt1[i] + beta[7]*ga1bwt2[i] +
        beta[8]*bwt1ga2[i] +
        beta[9]*male[i] + beta[10]*steroids[i] +
        u[unit[i]]
        }
```

```
    for (i in 1:10) {
        beta[i] ~ dnorm(0,0.00001)
        }
    for (j in 1:J) {
        u[j] ~ dnorm(beta[1], tau.u)
        du[j]<-u[j]-beta[1]
        }
    tau.u<-1/(sigma2.u)
    sigma2.u<-pow(sigma.u,2)
    sigma.u ~ dunif(0,5)
    }

#############################################################
#################          Model 2        ###################
#############################################################

model{
    for (i in 1:N) {
        died[i] ~ dbern(p[i])
        logit(p[i]) <- beta[2]*ga1[i] + beta[3]*ga2[i] +
        beta[4]*bwt1[i] +  beta[5]*bwt2[i] +
        beta[6]*ga1bwt1[i] + beta[7]*ga1bwt2[i] +
        beta[8]*bwt1ga2[i] +
        beta[9]*male[i] + beta[10]*steroids[i] +
        u[unit[i]]
        }
    for (i in 1:12) {
        beta[i] ~ dnorm(0,0.00001)
        }
    for (j in 1:J) {
        u[j] ~ dnorm(mu[j], tau.u)
        mu[j]<-beta[1] + beta[11]*level1[j] + beta[12]*level3[j]
        du[j]<-u[j]-mu[j]
        }
    tau.u<-1/(sigma2.u)
    sigma2.u<-pow(sigma.u,2)
    sigma.u ~ dunif(0,5)
    }

#############################################################
```

```
################          Model 3        ##################
#############################################################

model{
    for (i in 1:N) {
        died[i] ~ dbern(p[i])
        logit(p[i]) <- beta[2]*ga1[i] + beta[3]*ga2[i] +
        beta[4]*bwt1[i] +  beta[5]*bwt2[i] +
        beta[6]*ga1bwt1[i] + beta[7]*ga1bwt2[i] +
        beta[8]*bwt1ga2[i] +
        beta[9]*male[i] + beta[10]*steroids[i] +
        u[unit[i]]
        }
    for (i in 1:12) {
        beta[i] ~ dnorm(0,0.00001)
        }
    for (j in 1:J) {
        u[j] ~ dnorm(mu[j], tau.u[level[j]])
        mu[j]<-beta[1] + beta[11]*level1[j] + beta[12]*level3[j]
        du[j]<-u[j]-mu[j]
        }
    for (m in 1:3) {
        tau.u[m] <- 1/(sigma2.u[m])
        sigma2.u[m] <- pow(sigma.u[m], 2)
        sigma.u[m] ~ dunif(0, 5)
        }
    }

#############################################################
################          Model 4        ##################
#############################################################

model{
    for (i in 1:N) {
        died[i] ~ dbern(p[i])
        logit(p[i]) <- beta[2]*ga1[i] + beta[3]*ga2[i] +
        beta[4]*bwt1[i] +  beta[5]*bwt2[i] +
        beta[6]*ga1bwt1[i] + beta[7]*ga1bwt2[i] +
        beta[8]*bwt1ga2[i] +
        beta[9]*male[i] + beta[10]*steroids[i] +
```

```
        u[unit[i]]
        }
    for (i in 1:10) {
        beta[i] ~ dnorm(0,0.00001)
        }
    for (j in 1:J) {
        u[j] ~ dnorm(mu[j], tau.u)
        mu[j] <- v[network[j]]
        du[j]<-u[j]-mu[j]
        }
    for (k in 1:K) {
        v[k] ~ dnorm(beta[1], tau.v)
        dv[k] <- v[k] - beta[1]
        }
    tau.u<-1/(sigma2.u)
    sigma2.u<-pow(sigma.u,2)
    sigma.u ~ dunif(0,5)

    tau.v <- 1/(sigma2.v)
    sigma2.v <- pow(sigma.v, 2)
    sigma.v ~ dunif(0, 5)
}

#############################################################
#################          Model 5        ###################
#############################################################

model{
    for (i in 1:N) {
        died[i] ~ dbern(p[i])
        logit(p[i]) <- beta[2]*ga1[i] + beta[3]*ga2[i] +
        beta[4]*bwt1[i] +  beta[5]*bwt2[i] +
        beta[6]*ga1bwt1[i] + beta[7]*ga1bwt2[i] +
        beta[8]*bwt1ga2[i] +
        beta[9]*male[i] + beta[10]*steroids[i] +
        u[unit[i]]
        }
    for (i in 1:12) {
        beta[i] ~ dnorm(0,0.00001)
        }
```

```
    for (j in 1:J) {
        u[j] ~ dnorm(mu[j], tau.u)
        mu[j] <- v[network[j]] + beta[11]*level1[j] + beta[12]*level3[j]
        du[j]<-u[j]-mu[j]
        }
    for (k in 1:K) {
        v[k] ~ dnorm(beta[1], tau.v)
        dv[k] <- v[k] - beta[1]
        }
    tau.u<-1/(sigma2.u)
    sigma2.u<-pow(sigma.u,2)
    sigma.u ~ dunif(0,5)

    tau.v <- 1/(sigma2.v)
    sigma2.v <- pow(sigma.v, 2)
    sigma.v ~ dunif(0, 5)
}

############################################################
#################          Model 6        ##################
############################################################

model{
    for (i in 1:N) {
        died[i] ~ dbern(p[i])
        logit(p[i]) <- beta[2]*ga1[i] + beta[3]*ga2[i] +
        beta[4]*bwt1[i] +  beta[5]*bwt2[i] +
        beta[6]*ga1bwt1[i] + beta[7]*ga1bwt2[i] +
        beta[8]*bwt1ga2[i] +
        beta[9]*male[i] + beta[10]*steroids[i] +
        u[unit[i]]
        }
    for (i in 1:12) {
        beta[i] ~ dnorm(0,0.00001)
        }
    for (j in 1:J) {
        u[j] ~ dnorm(mu[j], tau.u[level[j]])
        mu[j] <- v[network[j]] + beta[11]*level1[j] + beta[12]*level3[j]
        du[j]<-u[j]-mu[j]
        }
```

```
    for (k in 1:K) {
        v[k] ~ dnorm(beta[1], tau.v)
        dv[k] <- v[k] - beta[1]
        }
    for (m in 1:3) {
        tau.u[m] <- 1/(sigma2.u[m])
        sigma2.u[m] <- pow(sigma.u[m], 2)
        sigma.u[m] ~ dunif(0, 5)
        }
    tau.v <- 1/(sigma2.v)
    sigma2.v <- pow(sigma.v, 2)
    sigma.v ~ dunif(0, 5)
}

#############################################################################
#############    Multiple membership (MM) model        ###################
#############################################################################

# wt[i,j] is the multiple membership weight assigned to unit[i,j]
    # the jth neonatal unit that the ith infant is in
# if an infant is cared for by t different neonatal units
    # then wt[i,j]=0 for all j>t
    # and unit[i,j]=170 for all j>t
    # where u[170] is assigned a value of 0
        # and does not contribute to the estimation of sigma.u
        # as J=169

model{
    for (i in 1:N) {
                died[i] ~ dbern(p[i])
                logit(p[i]) <- beta[2] * ga1[i] + beta[3] * ga2[i] +
        beta[4] * bwt1[i] + beta[5] * bwt2[i] + beta[6] *
        ga1bwt1[i] + beta[7] * ga1bwt2[i] + beta[8] * bwt1ga2[i] +
        beta[9] * male[i] + beta[10] * steroids[i] +
        wt[i,1] * u[unit[i,1]] + wt[i,2] * u[unit[i,2]] +
        wt[i,3] * u[unit[i,3]] + wt[i,4] * u[unit[i,4]] +
        wt[i,5] * u[unit[i,5]] + wt[i,6] * u[unit[i,6]] +
        wt[i,7] * u[unit[i,7]] + wt[i,8] * u[unit[i,8]] +
        wt[i,9] * u[unit[i,9]] + wt[i,10] * u[unit[i,10]] +
        wt[i,11] * u[unit[i,11]] + wt[i,12] * u[unit[i,12]]
```

```
        }
    u[170] <- 0
    for (i in 1:12) {
        beta[i] ~ dnorm(0, 0.00001)
        }
    for (j in 1:J) {
        u[j] ~ dnorm(mu[j], tau.u[j])
        mu[j] <- v[network[j]] + beta[11] * level1[j] + beta[12] * level3[j]
        du[j] <- u[j] - mu[j]
        tau.u[j] ~ dgamma(2, sigma2.u)
        }
    for (k in 1:K) {
        v[k] ~ dnorm(beta[1], tau.v)
        dv[k] <- v[k] - beta[1]
        }
    sigma2.u <- pow(sigma.u, 2)
    sigma.u ~ dunif(0, 5)

    tau.v <- 1/(sigma2.v)
    sigma2.v <- pow(sigma.v, 2)
    sigma.v ~ dunif(0, 5)
}
```

Appendix E

Additional results for chapter 6

E.1 Convergence statistics

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
Intercept	0.0081	0.0082	0.0101	0.0091	0.0087	0.0103
$GA^{(1)}$	0.0083	0.0084	0.0096	0.0091	0.0083	0.0097
$GA^{(2)}$	0.0077	0.0078	0.0088	0.0082	0.0078	0.0090
BWT	0.0020	0.0020	0.0021	0.0019	0.0018	0.0020
BWT^2	0.0009	0.0009	0.0009	0.0009	0.0008	0.0009
$GA^{(1)}\times$ BWT	0.0072	0.0074	0.0082	0.0076	0.0071	0.0082
$GA^{(1)}\times$ BWT^2	0.0012	0.0012	0.0013	0.0012	0.0011	0.0013
BWT $\times$ $GA^{(2)}$	0.0040	0.0041	0.0044	0.0040	0.0038	0.0043
Male	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004
Steroids	0.0011	0.0011	0.0011	0.0010	0.0009	0.0010
SCBU		0.0040	0.0042		0.0045	0.0025
NICU		0.0010	0.0010		0.0012	0.0009
σ_u	0.0016	0.0017		0.0022	0.0030	
$\sigma_{u_{SCBU}}$			0.0132			0.0056
$\sigma_{u_{LNU}}$			0.0055			0.0024
$\sigma_{u_{NICU}}$			0.0012			0.0013
σ_v				0.0014	0.0014	0.0015
u_2	0.0007	0.0007	0.0013	0.0005	0.0005	0.0010
u_{92}	0.0006	0.0006	0.0007	0.0005	0.0006	0.0006
u_{20}	0.0004	0.0005	0.0005	0.0004	0.0005	0.0008
u_{73}	0.0005	0.0004	0.0004	0.0005	0.0004	0.0007
u_{33}	0.0007	0.0007	0.0013	0.0005	0.0004	0.0010
u_{136}	0.0008	0.0010	0.0009	0.0010	0.0013	0.0012
u_{113}	0.0007	0.0007	0.0008	0.0006	0.0005	0.0004
u_{137}	0.0006	0.0006	0.0006	0.0005	0.0004	0.0004
u_{16}	0.0005	0.0005	0.0005	0.0004	0.0004	0.0003
u_{48}	0.0011	0.0009	0.0008	0.0014	0.0012	0.0012
v_{11}				0.0007	0.0009	0.0013
v_1				0.0023	0.0030	0.0032
v_{19}				0.0021	0.0023	0.0027
v_3				0.0006	0.0010	0.0014
v_7				0.0020	0.0022	0.0025

Table E.1: Monte Carlo standard errors (log scale), t-4 NNU effects U(0,5) prior

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
Intercept	865	878	515	677	706	566
$GA^{(1)}$	1291	1285	906	1031	1160	931
$GA^{(2)}$	1167	1128	818	962	1015	835
BWT	4507	4761	4232	4842	5710	4561
BWT^2	3976	4059	3768	4114	4972	4024
$GA^{(1)}\times$ BWT	1659	1601	1196	1416	1558	1250
$GA^{(1)}\times$ BWT^2	3376	3316	2824	3353	3733	3050
BWT $\times$ $GA^{(2)}$	2672	2630	2146	2584	2855	2308
Male	37706	39115	39303	44401	52727	44373
Steroids	10154	10116	9894	11335	13635	10860
SCBU		3942	4063		3181	13366
NICU		13451	14042		8642	15726
σ_u	2392	2172		1597	1062	
$\sigma_{u_{SCBU}}$			878			4175
$\sigma_{u_{LNU}}$			696			2381
$\sigma_{u_{NICU}}$			6835			6085
σ_v				4311	4336	3756
u_2	312781	296587	218904	395850	387862	400907
u_{92}	290541	257562	140939	358097	204439	115175
u_{20}	283826	204975	228915	275823	161207	98014
u_{73}	193038	284874	279508	147112	214424	99200
u_{33}	297684	286962	197225	389244	397215	394943
u_{136}	98818	70792	101813	59569	29812	43980
u_{113}	230422	246170	137244	280005	286649	273890
u_{137}	247345	273651	178654	265479	322956	252073
u_{16}	291009	282524	272352	391209	399583	403285
u_{48}	54920	66718	92157	28051	27645	38030
v_{11}				39095	27255	12779
v_1				9044	7026	5757
v_{19}				8008	7406	5397
v_3				45632	24568	11461
v_7				7808	6906	5802

Table E.2: Estimated sample size, t-4 NNU effects U(0,5) prior

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
Intercept	0.0078	0.0077	0.0105	0.0094	0.0096	0.0118
$GA^{(1)}$	0.0078	0.0081	0.0104	0.0097	0.0095	0.0109
$GA^{(2)}$	0.0073	0.0074	0.0096	0.0088	0.0088	0.0102
BWT	0.0021	0.0021	0.0024	0.0022	0.0021	0.0022
BWT^2	0.0009	0.0010	0.0011	0.0010	0.0010	0.0010
$GA^{(1)}\times$ BWT	0.0069	0.0070	0.0091	0.0081	0.0082	0.0091
$GA^{(1)}\times$ BWT^2	0.0012	0.0012	0.0015	0.0013	0.0013	0.0014
BWT $\times$ $GA^{(2)}$	0.0040	0.0040	0.0050	0.0045	0.0044	0.0047
Male	0.0005	0.0005	0.0005	0.0004	0.0004	0.0004
Steroids	0.0012	0.0011	0.0013	0.0012	0.0011	0.0011
SCBU		0.0038	0.0037		0.0042	0.0028
NICU		0.0009	0.0009		0.0011	0.0010
σ_u	0.0006	0.0007		0.0013	0.0023	
$\sigma_{u_{SCBU}}$			0.0073			0.0066
$\sigma_{u_{LNU}}$			0.0035			0.0028
$\sigma_{u_{NICU}}$			0.0005			0.0016
σ_v				0.0014	0.0015	0.0018
u_2	0.0008	0.0008	0.0013	0.0005	0.0005	0.0012
u_{92}	0.0007	0.0007	0.0008	0.0005	0.0006	0.0007
u_{20}	0.0006	0.0006	0.0006	0.0005	0.0006	0.0009
u_{73}	0.0005	0.0005	0.0005	0.0006	0.0005	0.0008
u_{33}	0.0008	0.0008	0.0014	0.0006	0.0005	0.0012
u_{136}	0.0007	0.0008	0.0007	0.0010	0.0014	0.0014
u_{113}	0.0008	0.0008	0.0007	0.0005	0.0005	0.0005
u_{137}	0.0007	0.0007	0.0006	0.0006	0.0005	0.0005
u_{16}	0.0007	0.0007	0.0005	0.0005	0.0004	0.0004
u_{48}	0.0008	0.0007	0.0007	0.0013	0.0014	0.0014
v_{11}				0.0006	0.0008	0.0015
v_1				0.0025	0.0031	0.0038
v_{19}				0.0020	0.0025	0.0031
v_3				0.0006	0.0011	0.0015
v_7				0.0020	0.0023	0.0029

Table E.3: Monte Carlo standard errors (log scale), Normal NNU effects U(0,5) prior

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
Intercept	878	937	503	664	637	429
$GA^{(1)}$	1358	1274	787	943	948	728
$GA^{(2)}$	1228	1173	707	879	855	639
BWT	4290	4186	3122	3689	4059	3724
BWT^2	3918	3712	2693	3342	3577	3290
$GA^{(1)}\times$ BWT	1688	1640	1006	1280	1236	992
$GA^{(1)}\times$ BWT^2	3171	3112	2198	2741	2772	2487
BWT $\times$ $GA^{(2)}$	2625	2569	1690	2094	2159	1873
Male	30246	30528	26295	34599	36194	34763
Steroids	8519	9114	6896	8727	9554	8889
SCBU		4634	5692		3662	10408
NICU		15448	16122		10203	11625
σ_u	9878	8515		3595	1692	
$\sigma_{u_{SCBU}}$			2209			3035
$\sigma_{u_{LNU}}$			1445			1849
$\sigma_{u_{NICU}}$			29217			4521
σ_v				4237	4122	2846
u_2	199554	200000	200444	300000	301398	297911
u_{92}	200000	193033	104718	296965	169643	88661
u_{20}	191586	195101	197967	240791	135528	73262
u_{73}	179632	197951	198701	122319	182185	75061
u_{33}	200924	200000	184253	295539	298882	298990
u_{136}	142010	112736	157502	61953	26201	33248
u_{113}	196879	197542	159454	292066	281036	216328
u_{137}	191764	198989	147763	257337	261345	187555
u_{16}	200000	200000	204813	309626	309894	300695
u_{48}	99161	104933	152402	33126	23128	28133
v_{11}				49605	28511	9476
v_1				8185	6188	4154
v_{19}				8607	6453	4159
v_3				47745	21032	9676
v_7				7716	6363	4228

Table E.4: Estimated sample size, Normal NNU effects U(0,5) prior

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
Intercept	0.0083	0.0081	0.0107	0.0086	0.0098	0.0124
$GA^{(1)}$	0.0089	0.0085	0.0107	0.0088	0.0098	0.0115
$GA^{(2)}$	0.0082	0.0079	0.0099	0.0080	0.0091	0.0107
BWT	0.0022	0.0022	0.0024	0.0020	0.0021	0.0023
BWT^2	0.0010	0.0010	0.0011	0.0009	0.0010	0.0011
$GA^{(1)}\times$ BWT	0.0076	0.0075	0.0092	0.0075	0.0083	0.0097
$GA^{(1)}\times$ BWT^2	0.0013	0.0013	0.0015	0.0012	0.0013	0.0015
BWT $\times$ $GA^{(2)}$	0.0044	0.0043	0.0050	0.0041	0.0045	0.0050
Male	0.0005	0.0005	0.0005	0.0004	0.0004	0.0004
Steroids	0.0012	0.0012	0.0013	0.0010	0.0011	0.0012
SCBU		0.0039	0.0040		0.0044	0.0033
NICU		0.0010	0.0009		0.0011	0.0010
σ_u	0.0006	0.0007		0.0014	0.0021	
$\sigma_{u_{SCBU}}$			0.0079			0.0064
$\sigma_{u_{LNU}}$			0.0032			0.0030
$\sigma_{u_{NICU}}$			0.0005			0.0015
σ_v				0.0015	0.0014	0.0019
u_2	0.0008	0.0008	0.0014	0.0005	0.0005	0.0012
u_{92}	0.0007	0.0007	0.0008	0.0005	0.0006	0.0007
u_{20}	0.0006	0.0006	0.0006	0.0005	0.0006	0.0009
u_{73}	0.0005	0.0005	0.0005	0.0006	0.0005	0.0008
u_{33}	0.0008	0.0008	0.0013	0.0006	0.0005	0.0012
u_{136}	0.0007	0.0008	0.0007	0.0010	0.0014	0.0014
u_{113}	0.0008	0.0008	0.0007	0.0005	0.0005	0.0004
u_{137}	0.0007	0.0007	0.0006	0.0006	0.0005	0.0004
u_{16}	0.0007	0.0007	0.0005	0.0005	0.0004	0.0003
u_{48}	0.0008	0.0008	0.0007	0.0014	0.0014	0.0014
v_{11}				0.0007	0.0009	0.0017
v_1				0.0026	0.0032	0.0040
v_{19}				0.0021	0.0023	0.0034
v_3				0.0006	0.0011	0.0017
v_7				0.0022	0.0022	0.0032

Table E.5: Monte Carlo standard errors (log scale), Normal NNU effects U(0,100) prior

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
Intercept	822	885	504	719	637	401
$GA^{(1)}$	1115	1216	765	1075	923	678
$GA^{(2)}$	1033	1079	693	989	811	603
BWT	3680	3759	3158	4547	4144	3451
BWT^2	3342	3389	2694	3937	3482	3118
$GA^{(1)}\times$ BWT	1458	1472	998	1425	1219	925
$GA^{(1)}\times$ BWT^2	2751	2843	2077	3233	2706	2316
BWT $\times$ $GA^{(2)}$	2234	2271	1674	2503	2140	1731
Male	25907	27542	26670	39137	34753	32919
Steroids	7574	7830	6929	10526	9310	8746
SCBU		4469	4746		3511	7628
NICU		14869	14968		10025	13034
σ_u	9540	8483		3502	1849	
$\sigma_{u_{SCBU}}$			1949			3258
$\sigma_{u_{LNU}}$			1578			1603
$\sigma_{u_{NICU}}$			27644			4528
σ_v				3804	4424	2495
u_2	200000	200000	187926	300000	302722	289837
u_{92}	197768	197885	110166	269540	182215	93709
u_{20}	193948	189542	193367	229149	138555	74440
u_{73}	175922	195614	198939	116461	186212	71457
u_{33}	199246	200000	192154	297456	296303	294024
u_{136}	146421	115517	158498	54303	27863	30956
u_{113}	196856	200000	137589	296778	267938	241452
u_{137}	197265	190255	154809	254033	281820	214390
u_{16}	200000	196258	199290	305941	301670	304090
u_{48}	96096	105241	146491	30497	24942	28964
v_{11}				42256	26321	7612
v_1				7172	6266	3683
v_{19}				7673	7147	3439
v_3				47587	22076	8023
v_7				6627	7052	3561

Table E.6: Estimated sample size, Normal NNU effects U(0,100) prior

E.2 Sensitivity analysis with alternative assumptions for NNU effects

E.2.1 Normal distribution for NNU effects

Table E.7: Posterior mean odds ratios and 95% credible intervals for infant and provider characteristics, Normal NNU effects U(0,5) prior

	Model 1		Model 2		Model 3		Model 4		Model 5		Model 6	
	OR	95% CrI	OR	95% CrI	OR	95% CrI	OR	95% CrI	OR	95% CrI	OR	95% CrI
$GA^{(1)}$	0.09	(0.05,0.15)	0.09	(0.05,0.16)	0.09	(0.05,0.16)	0.08	(0.05,0.15)	0.09	(0.05,0.16)	0.09	(0.05,0.16)
$GA^{(2)}$	3.05	(1.85,5.02)	2.94	(1.77,4.81)	2.88	(1.75,4.79)	3.14	(1.88,5.24)	3.02	(1.81,4.94)	2.96	(1.80,4.96)
BWT	0.87	(0.67,1.13)	0.88	(0.68,1.14)	0.89	(0.68,1.15)	0.86	(0.66,1.12)	0.87	(0.67,1.14)	0.88	(0.67,1.15)
BWT^2	2.03	(1.81,2.28)	2.02	(1.80,2.26)	2.02	(1.80,2.27)	2.04	(1.82,2.29)	2.02	(1.80,2.27)	2.02	(1.80,2.27)
$GA^{(1)}\times$ BWT	0.27	(0.15,0.47)	0.27	(0.16,0.48)	0.28	(0.16,0.49)	0.26	(0.15,0.46)	0.27	(0.15,0.47)	0.27	(0.15,0.48)
$GA^{(1)}\times BWT^2$	0.78	(0.68,0.89)	0.78	(0.68,0.89)	0.78	(0.68,0.90)	0.78	(0.68,0.89)	0.78	(0.68,0.89)	0.78	(0.68,0.89)
BWT $\times$ $GA^{(2)}$	1.31	(0.88,1.95)	1.30	(0.87,1.92)	1.28	(0.86,1.92)	1.33	(0.88,1.99)	1.31	(0.87,1.96)	1.29	(0.87,1.94)
Male	1.60	(1.36,1.87)	1.60	(1.36,1.87)	1.59	(1.36,1.87)	1.60	(1.36,1.88)	1.60	(1.36,1.87)	1.60	(1.36,1.87)
Steroids	0.50	(0.41,0.62)	0.50	(0.40,0.62)	0.50	(0.41,0.62)	0.50	(0.41,0.62)	0.50	(0.41,0.62)	0.50	(0.40,0.62)
SCBU	-	-	1.52	(0.91,2.48)	1.47	(0.83,2.48)	-	-	1.45	(0.88,2.38)	1.44	(0.80,2.46)
NICU	-	-	1.28	(1.02,1.61)	1.26	(1.01,1.59)	-	-	1.32	(1.06,1.64)	1.31	(1.05,1.64)

OR Odds ratio; CrI Credible interval

$GA^{(1)}$ and $GA^{(2)}$ denote spline terms for gestational age at birth; BWT birth weight

All continuous coefficients were standardised so OR are for a one standard deviation increase

* SCBU odds ratio comparing special care baby units to local neonatal units (LNUs); NICU odds ratio comparing neonatal intensive care units to LNUs

	DIC	Deviance	Effective parameters
Model 1	4728.00	4676.00	52.37
Model 2	4727.00	4675.00	51.97
Model 3	4728.00	4677.00	51.76
Model 4	4729.00	4678.00	50.66
Model 5	4726.00	4679.00	47.68
Model 6	4726.00	4675.00	51.62

Table E.8: Model performance, Normal NNU effects U(0,5) prior

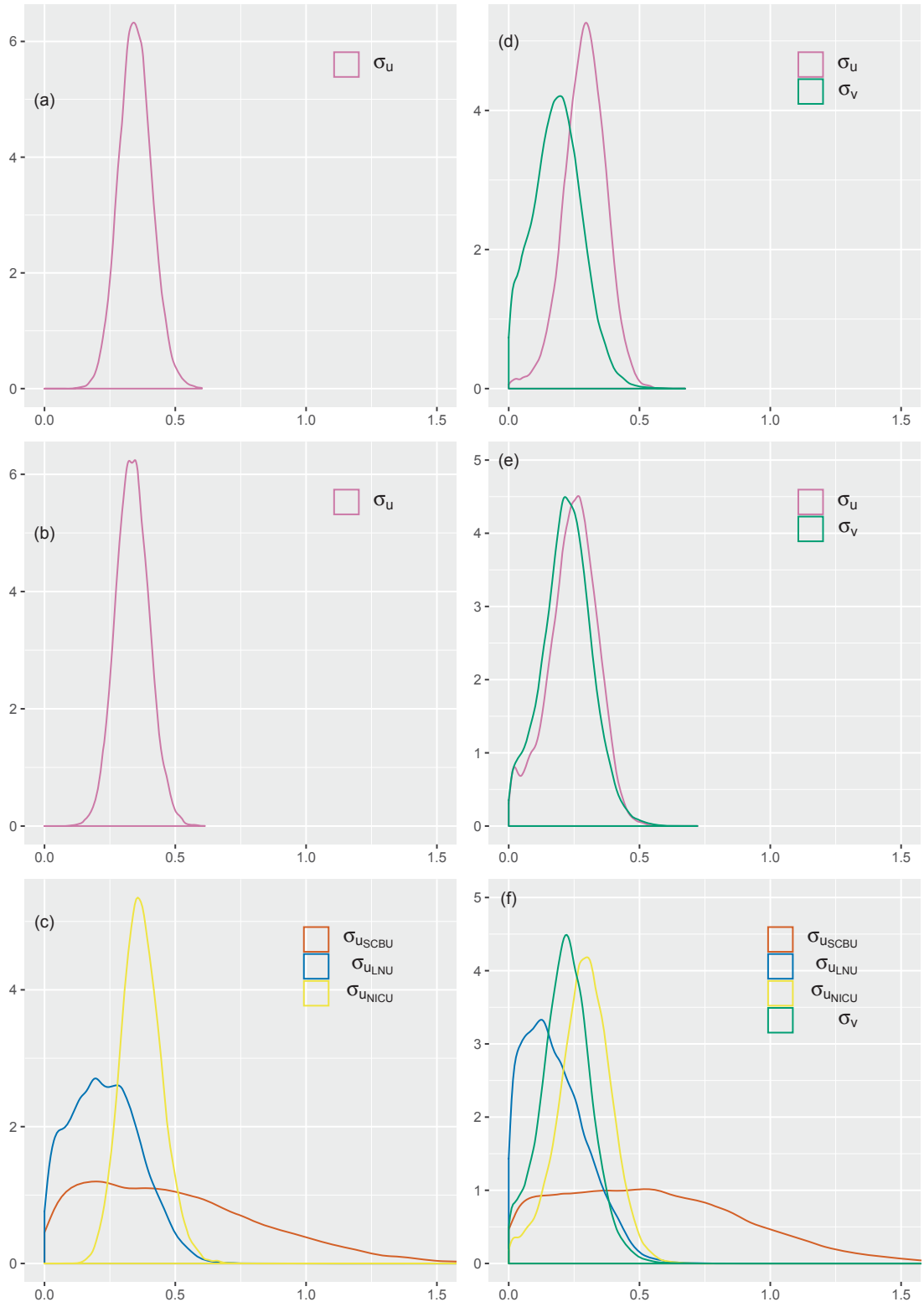


Figure E.1: Posterior distribution provider SD terms, Normal NNU effects, U(0,5) prior for σ_u. Models 1-6 shown in figures (a) to (f)

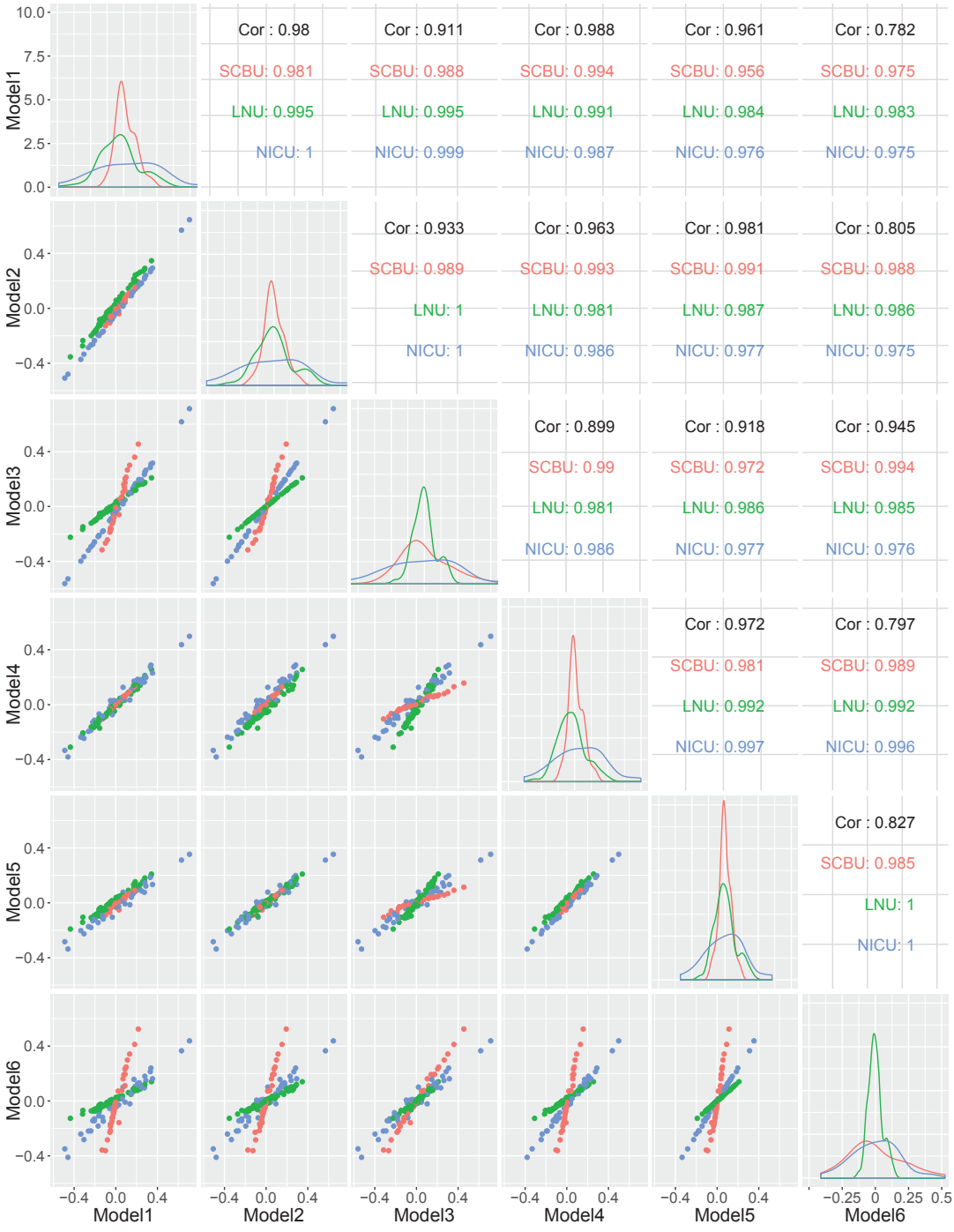


Figure E.2: Correlation of NNU effects across models, Normal distribution for NNU effects with U(0,5) prior for σ_u

E.2.2 Caterpillar plots for Normal NNU effects

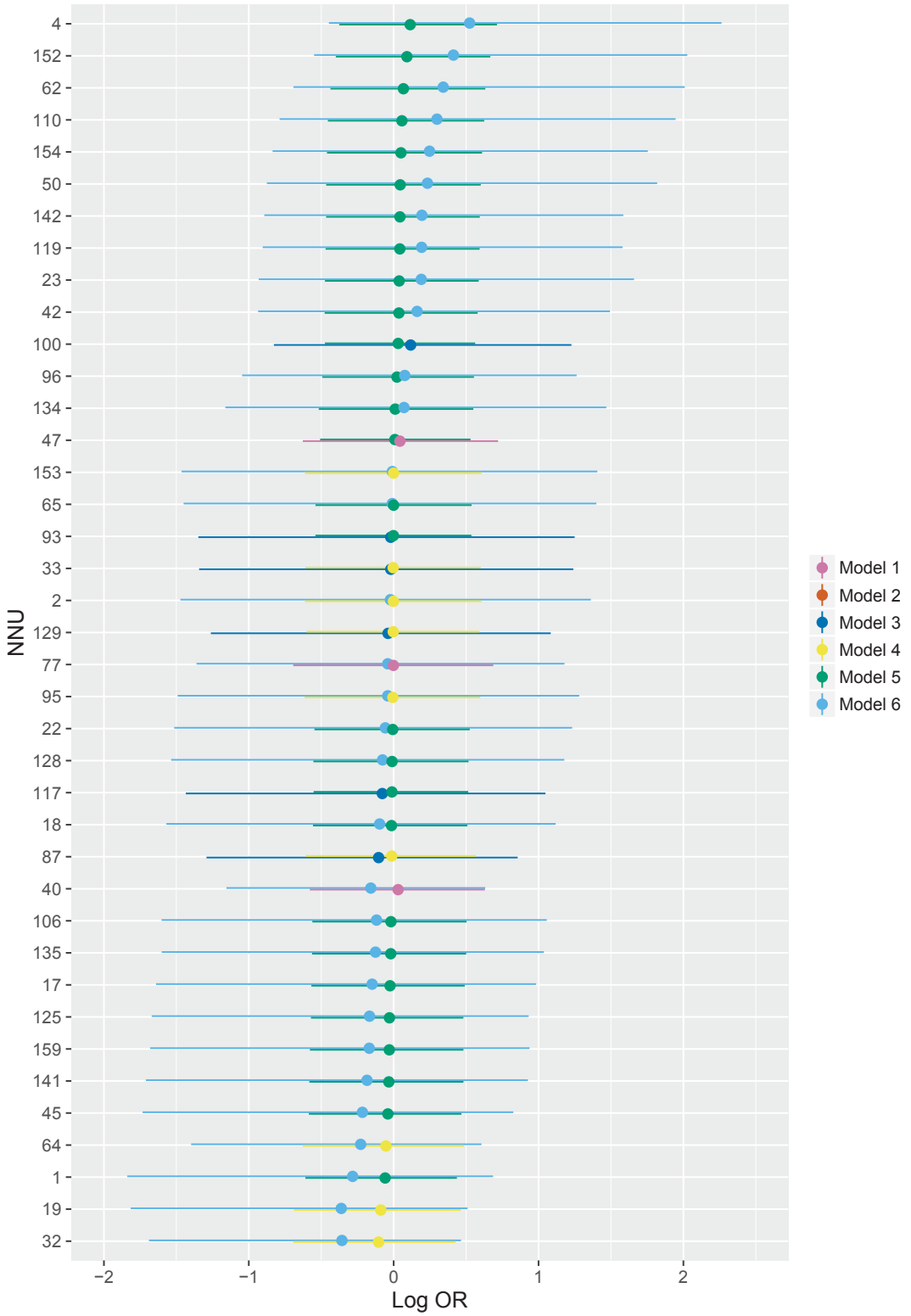


Figure E.3: Caterpillar plot for SCBU, showing smallest and largest effects across models, Normal NNU effects

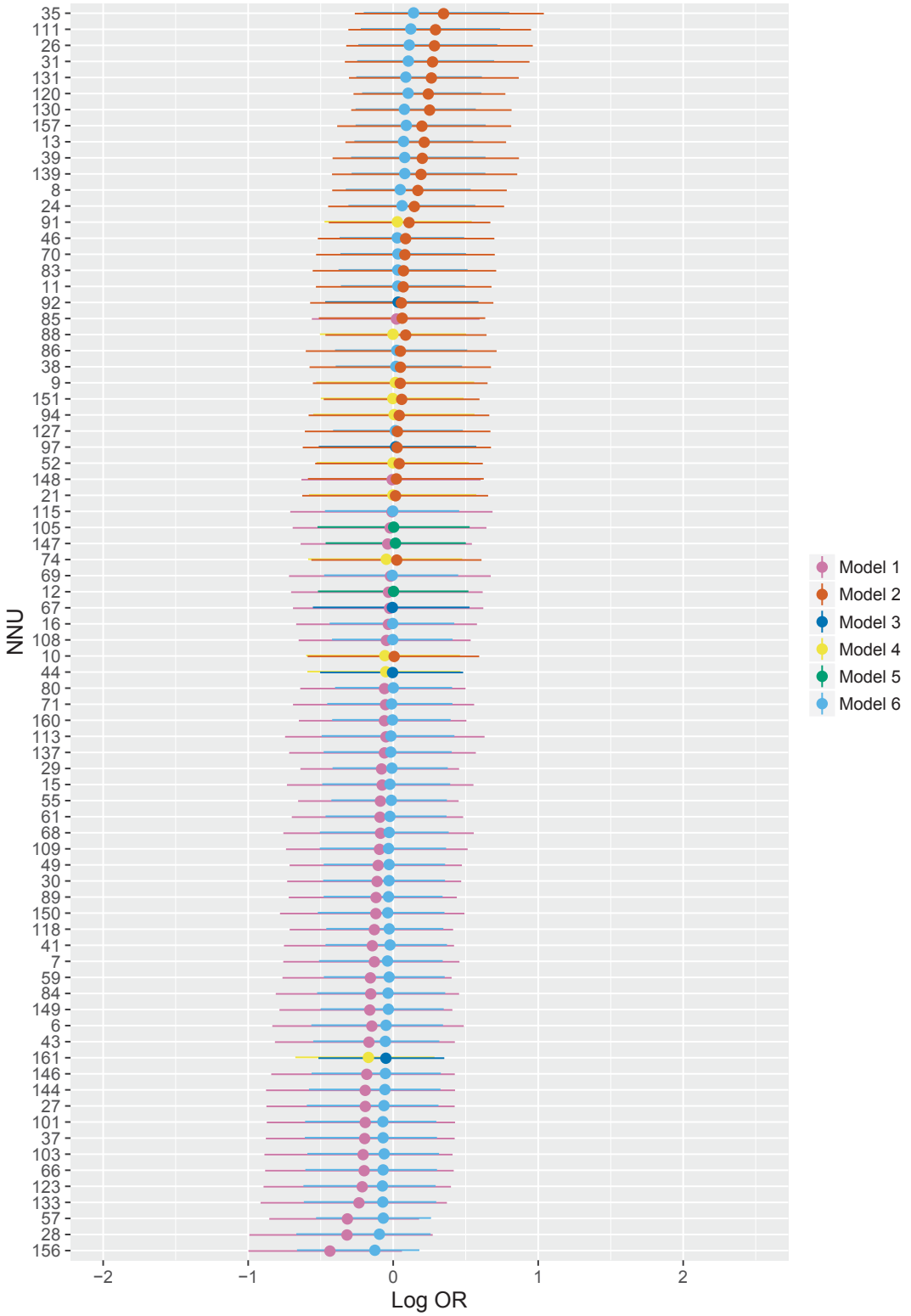


Figure E.4: Caterpillar plot for LNU, showing smallest and largest effects across models, Normal NNU effects

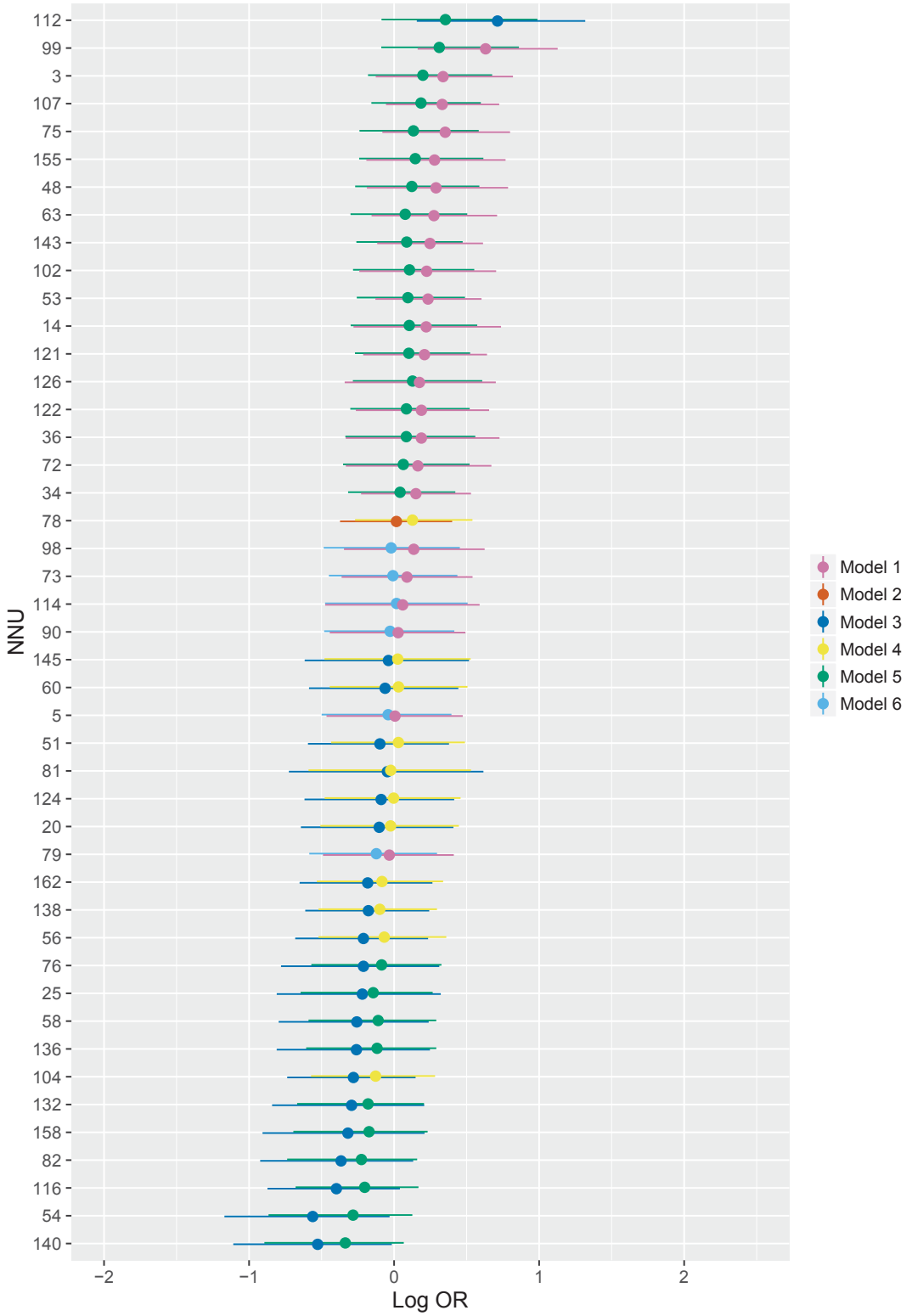


Figure E.5: Caterpillar plot for NICU, showing smallest and largest effects across models, Normal NNU effects

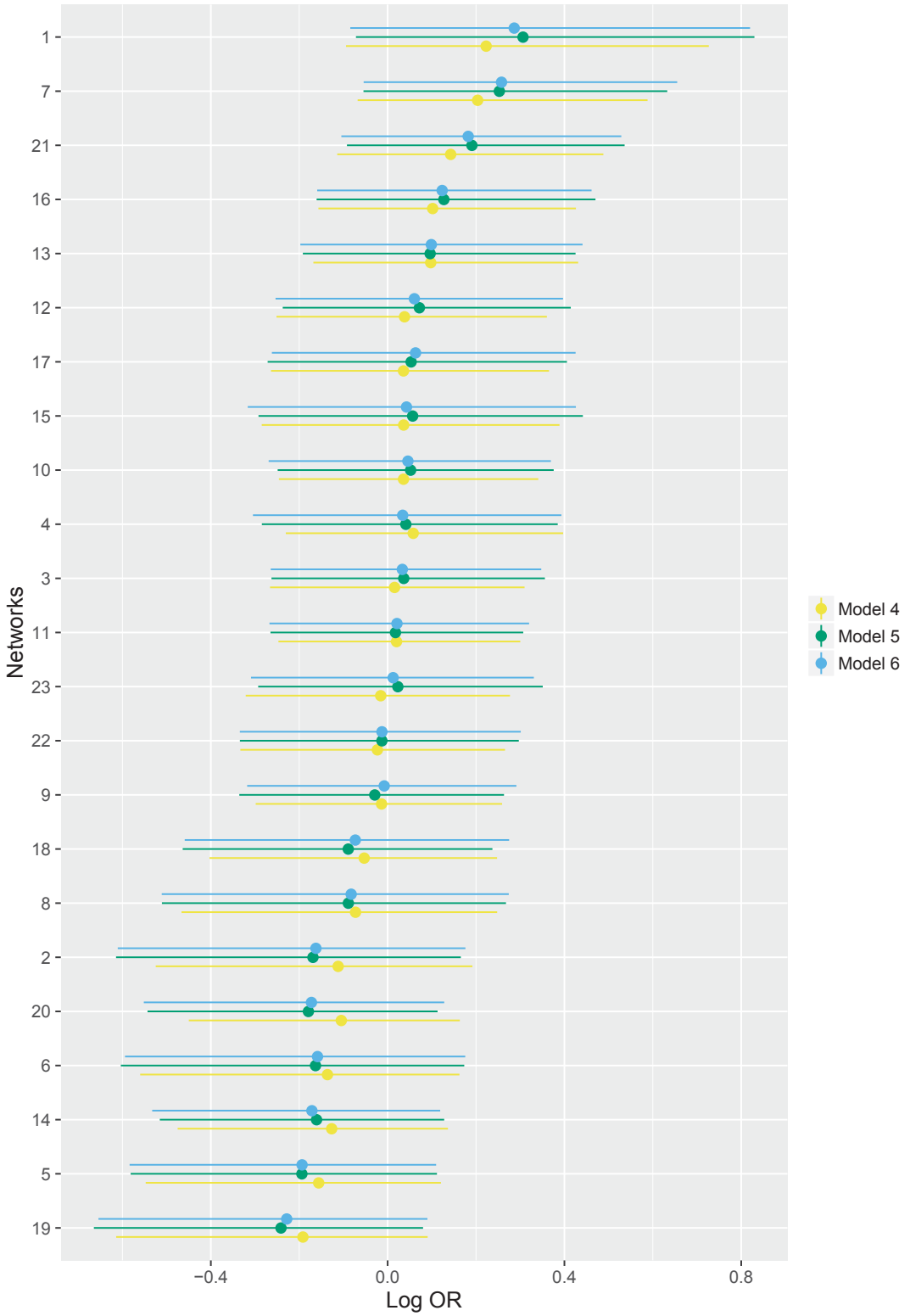


Figure E.6: Caterpillar plot for networks, showing smallest and largest effects across models, Normal NNU effects

E.2.3 U(0,100) prior for the standard deviation of NNU effects

	DIC	Deviance	Effective parameters
Model 1	4728.00	4676.00	52.38
Model 2	4727.00	4675.00	52.17
Model 3	4728.00	4677.00	51.71
Model 4	4729.00	4678.00	50.54
Model 5	4726.00	4678.00	47.88
Model 6	4726.00	4675.00	51.24

Table E.9: Model performance, Normal NNU effects U(0,100) prior

Table E.10: Posterior mean odds ratios and 95% credible intervals for infant and provider characteristics, Normal NNU effects U(0,5) prior

	Model 1		Model 2		Model 3		Model 4		Model 5		Model 6	
	Mean	95% CI	Mean	95% CI	Mean	95% CI	Mean	95% CI	Mean	95% CI	Mean	95% CI
$GA^{(1)}$	0.09	(0.05,0.15)	0.09	(0.05,0.16)	0.09	(0.05,0.16)	0.09	(0.05,0.15)	0.09	(0.05,0.16)	0.09	(0.05,0.16)
$GA^{(2)}$	3.08	(1.84,5.11)	2.98	(1.79,4.96)	2.91	(1.76,4.89)	3.09	(1.89,5.05)	3.01	(1.82,5.02)	2.95	(1.77,4.90)
BWT	0.87	(0.66,1.13)	0.88	(0.67,1.14)	0.88	(0.68,1.15)	0.87	(0.66,1.13)	0.87	(0.67,1.14)	0.88	(0.68,1.15)
BWT^2	2.03	(1.81,2.28)	2.03	(1.80,2.27)	2.02	(1.80,2.27)	2.04	(1.82,2.28)	2.02	(1.80,2.27)	2.02	(1.80,2.27)
$GA^{(1)}\times$ BWT	0.27	(0.15,0.47)	0.27	(0.15,0.48)	0.28	(0.16,0.49)	0.27	(0.15,0.46)	0.27	(0.15,0.47)	0.27	(0.16,0.48)
$GA^{(1)}\times BWT^2$	0.78	(0.68,0.89)	0.78	(0.68,0.89)	0.78	(0.68,0.90)	0.78	(0.68,0.89)	0.78	(0.68,0.89)	0.78	(0.68,0.90)
BWT $\times$ $GA^{(2)}$	1.32	(0.88,1.97)	1.30	(0.87,1.95)	1.29	(0.86,1.93)	1.31	(0.88,1.96)	1.31	(0.88,1.96)	1.29	(0.86,1.93)
Male	1.60	(1.36,1.87)	1.60	(1.36,1.87)	1.60	(1.36,1.87)	1.60	(1.36,1.87)	1.60	(1.36,1.87)	1.60	(1.36,1.87)
Steroids	0.50	(0.41,0.62)	0.50	(0.41,0.62)	0.50	(0.41,0.62)	0.50	(0.41,0.62)	0.50	(0.40,0.62)	0.50	(0.40,0.62)

OR Odds ratio; CrI Credible interval

$GA^{(1)}$ and $GA^{(2)}$ denote spline terms for gestational age at birth; BWT birth weight

All continuous coefficients were standardised so OR are for a one standard deviation increase

* SCBU odds ratio comparing special care baby units to local neonatal units (LNUs); NICU odds ratio comparing neonatal intensive care units to LNUs

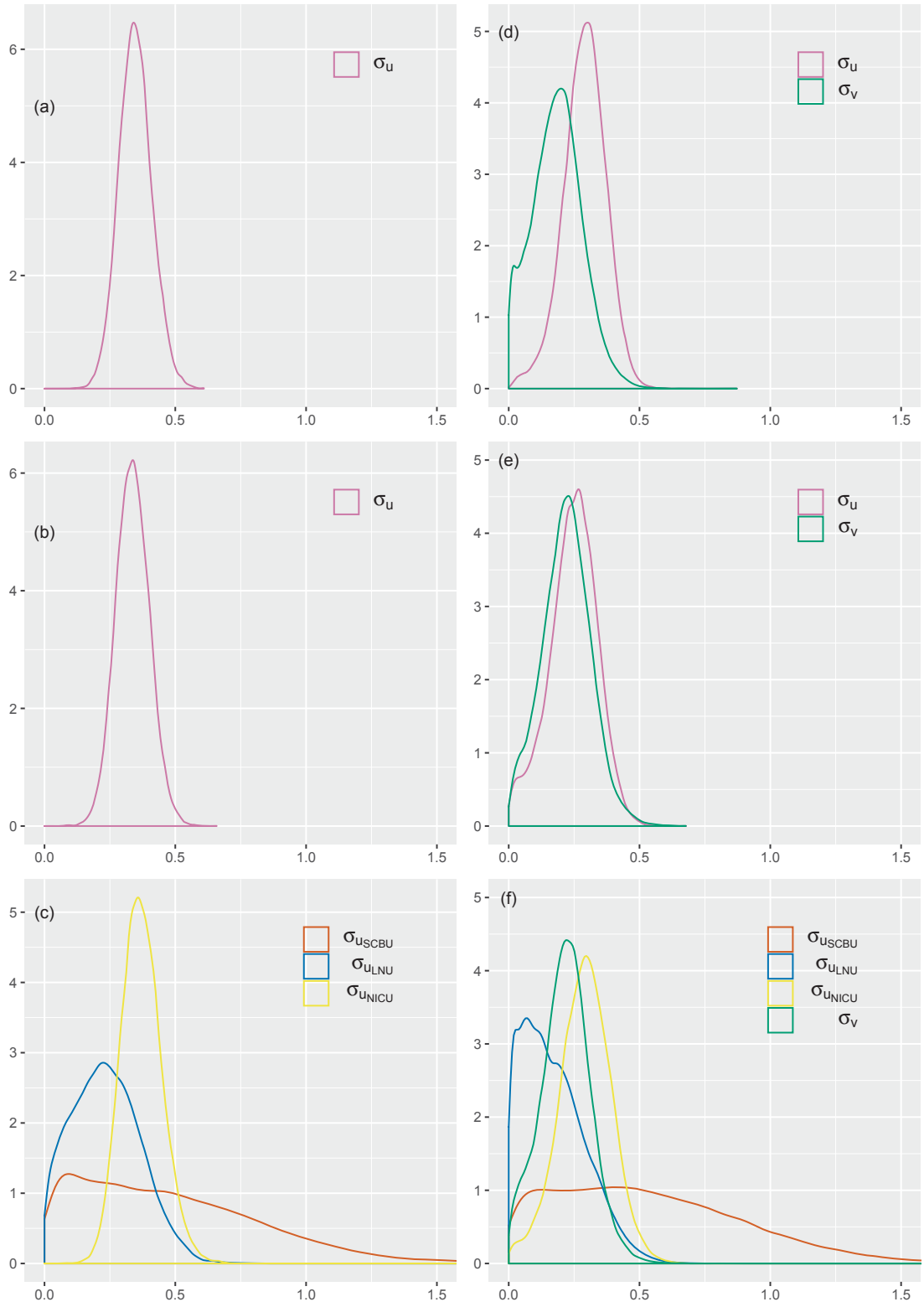


Figure E.7: Posterior distribution provider SD terms, Normal NNU effects, U(0,100) prior for σ_u. Models 1-6 shown in figures (a) to (f)

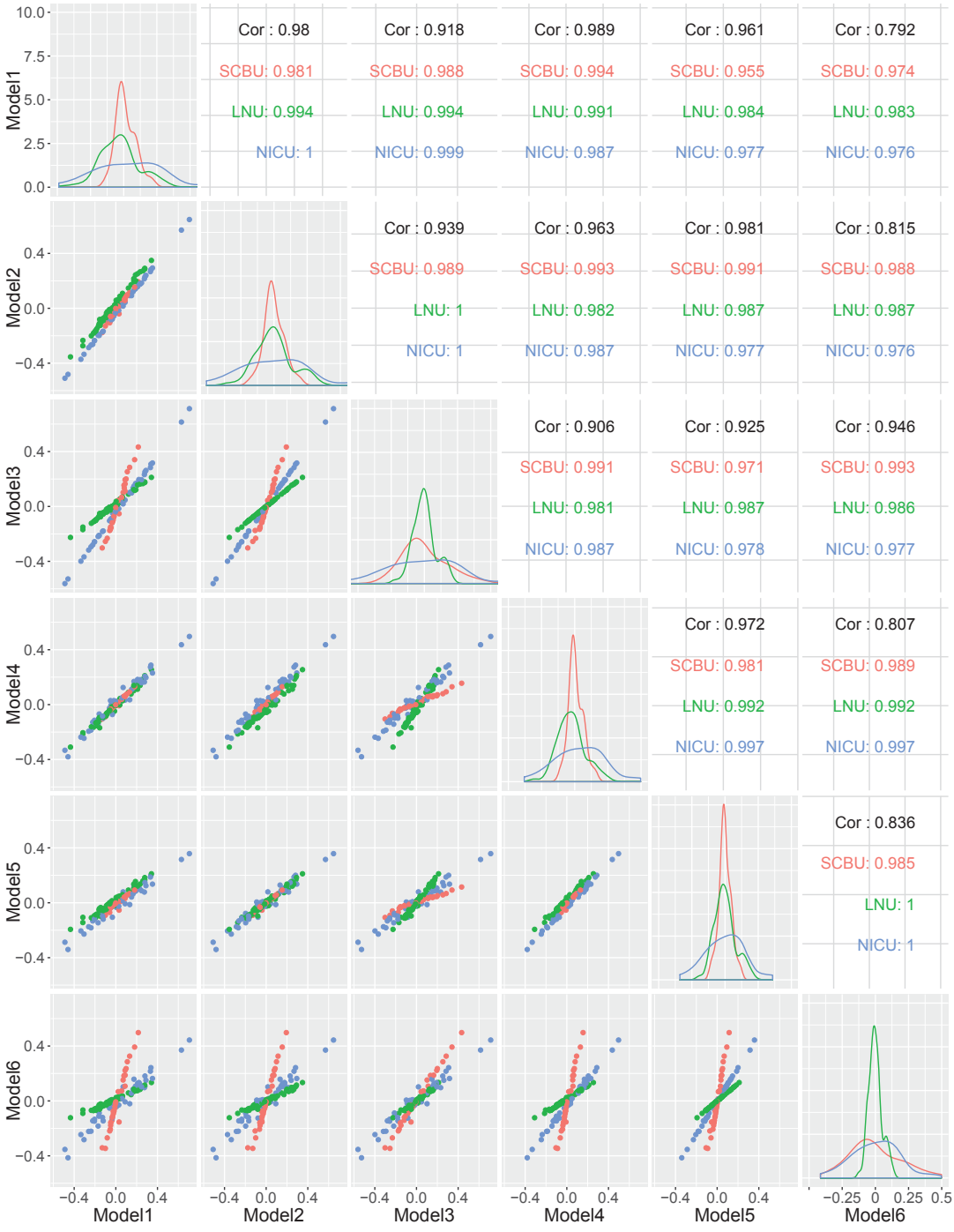


Figure E.8: Correlation of NNU effects across models, Normal distribution for NNU effects with U(0,100) prior for σ_u

Appendix F

Additional results for chapter 7

F.1 Convergence statistics

	Equal	*LOS*	*Beta2*	*Beta3*	*WLOS*	*First*
Intercept	0.0064	0.0063	0.0080	0.0076	0.0064	0.0091
$GA^{(1)}$	0.0067	0.0066	0.0082	0.0080	0.0067	0.0093
$GA^{(2)}$	0.0063	0.0061	0.0078	0.0076	0.0062	0.0089
BWT	0.0013	0.0013	0.0015	0.0014	0.0013	0.0014
BWT^2	9e-04	9e-04	0.0011	0.0010	9e-04	0.0011
$GA^{(1)}\times$ BWT	0.0061	0.0060	0.0074	0.0072	0.0061	0.0080
$GA^{(1)}\times BWT^2$	0.0010	0.0010	0.0012	0.0011	0.0010	0.0012
BWT $\times$ $GA^{(2)}$	0.0034	0.0034	0.0041	0.0038	0.0034	0.0040
Male	2e-04	3e-04	3e-04	2e-04	2e-04	3e-04
Steroids	6e-04	6e-04	7e-04	6e-04	6e-04	6e-04
SCBU	0.0036	0.0040	0.0051	0.0055	0.0051	0.0029
NICU	0.0011	0.0010	0.0011	0.0013	0.0011	0.0011
σ_u	0.0026	0.0022	0.0025	0.0034	0.0020	0.0029
σ_v	7e-04	7e-04	6e-04	7e-04	6e-04	7e-04
u_2	7e-04	7e-04	4e-04	4e-04	5e-04	5e-04
u_{92}	6e-04	8e-04	6e-04	5e-04	6e-04	3e-04
u_{20}	0.0010	9e-04	7e-04	8e-04	9e-04	4e-04
u_{73}	4e-04	5e-04	3e-04	3e-04	4e-04	2e-04
u_{33}	4e-04	6e-04	5e-04	6e-04	6e-04	3e-04
u_{136}	4e-04	5e-04	4e-04	4e-04	4e-04	4e-04
u_{113}	5e-04	5e-04	3e-04	3e-04	4e-04	3e-04
u_{137}	0.0015	0.0015	0.0012	0.0012	0.0013	0.0011
u_{16}	6e-04	7e-04	6e-04	5e-04	6e-04	3e-04
u_{48}	6e-04	6e-04	7e-04	9e-04	5e-04	4e-04
v_{11}	7e-04	7e-04	8e-04	0.0010	7e-04	0.0012
v_1	9e-04	8e-04	0.0010	0.0013	8e-04	0.0013
v_{19}	0.0012	0.0011	0.0013	0.0015	0.0012	0.0016
v_3	9e-04	8e-04	9e-04	0.0011	8e-04	0.0012
v_7	0.0010	9e-04	9e-04	9e-04	8e-04	0.0011

Table F.1: Monte Carlo standard errors (log scale) for multiple membership models

	Equal	*LOS*	*Beta2*	*Beta3*	*WLOS*	*First*
Intercept	955	1066	679	619	944	412
$GA^{(1)}$	1380	1531	1031	985	1422	692
$GA^{(2)}$	1218	1374	887	825	1271	587
BWT	7676	7341	5921	6512	7396	6235
BWT^2	3762	3864	2836	3079	3788	2648
$GA^{(1)}\times$ BWT	1822	1978	1337	1315	1854	1037
$GA^{(1)}\times BWT^2$	3829	3851	2802	3075	3734	2755
BWT $\times$ $GA^{(2)}$	3021	3097	2227	2333	2963	2126
Male	64569	62972	58166	63755	68063	57866
Steroids	18654	16929	13618	15438	17765	15723
SCBU	3458	3804	2147	1652	2516	2160
NICU	8703	11945	8279	5574	9652	5139
σ_u	1137	1580	981	639	1526	577
σ_v	9524	12129	13605	9607	14159	6309
u_2	181593	236928	307950	317821	271984	94099
u_{92}	229044	170276	168431	183368	208445	341133
u_{20}	69312	104431	100132	64542	96170	122916
u_{73}	322773	295824	329010	335330	317608	346927
u_{33}	270971	206875	177845	108229	181442	194784
u_{136}	224215	251526	266534	194300	250006	85895
u_{113}	202599	298810	341805	350129	320472	277596
u_{137}	26853	33833	31374	22738	36045	18172
u_{16}	177999	167622	133561	130977	175351	195291
u_{48}	132384	182358	77586	42446	205097	96233
v_{11}	38524	44563	28093	17673	38146	9410
v_1	34177	49453	29150	16206	40590	12276
v_{19}	18868	24588	17180	11004	21531	7457
v_3	27069	36982	23911	13847	31185	9394
v_7	21465	29860	27849	21211	33426	12137

Table F.2: Estimated sample size for multiple membership models

F.2 NNU effects ordered by proportion transferred

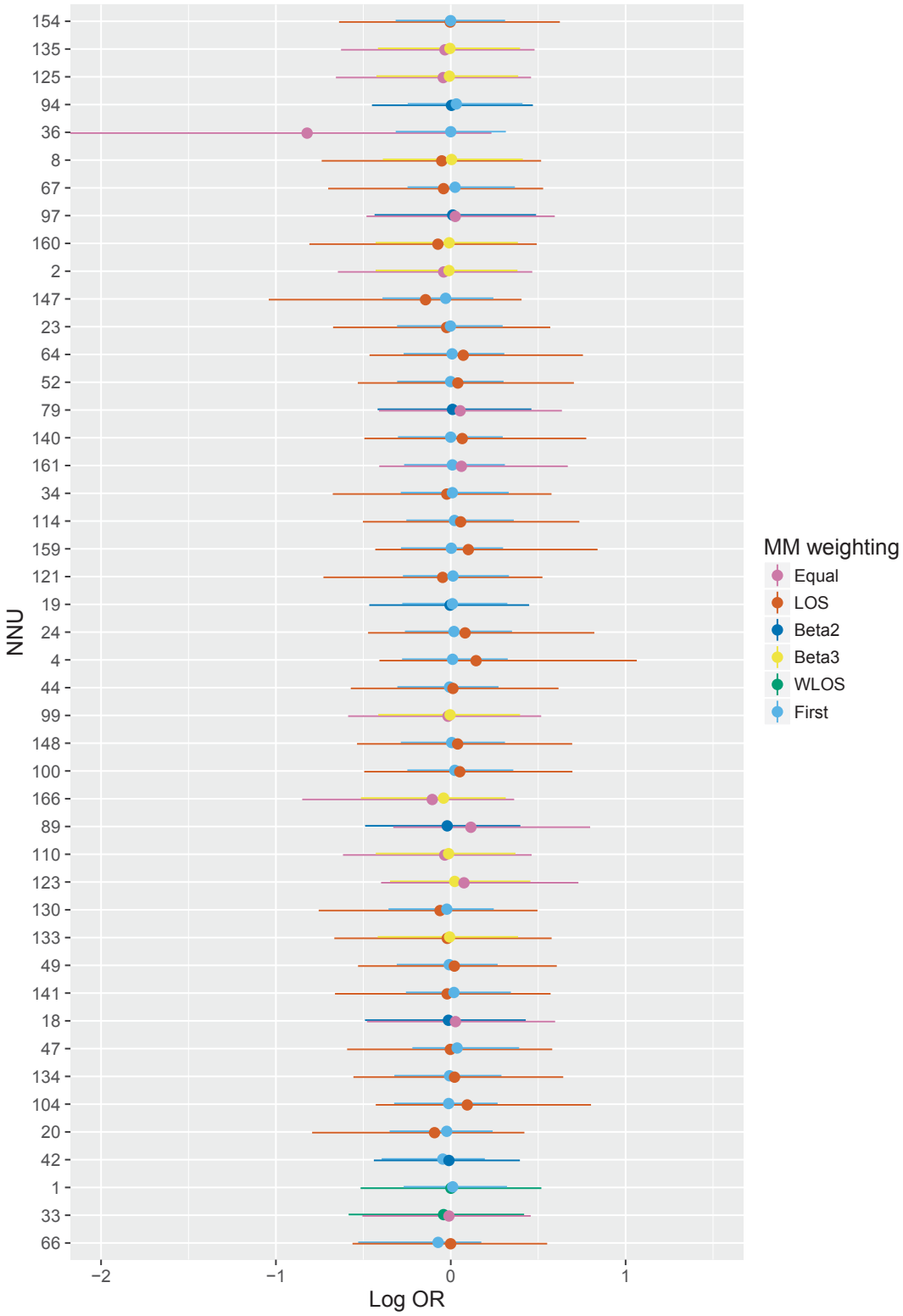


Figure F.1: Caterpillar plots for SCBU effects, ordered by proportion transferred

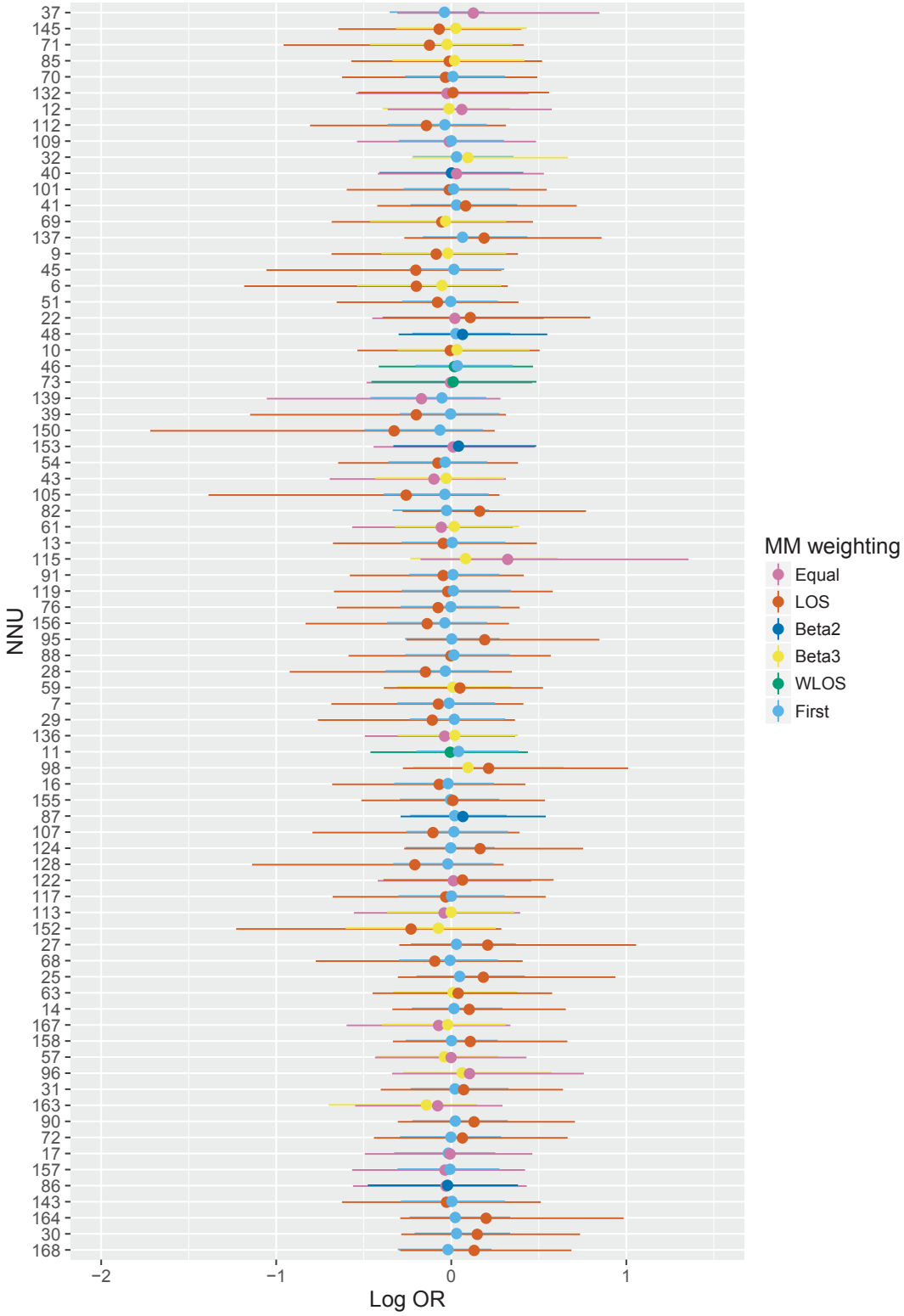


Figure F.2: Caterpillar plots for LNU effects, ordered by proportion transferred

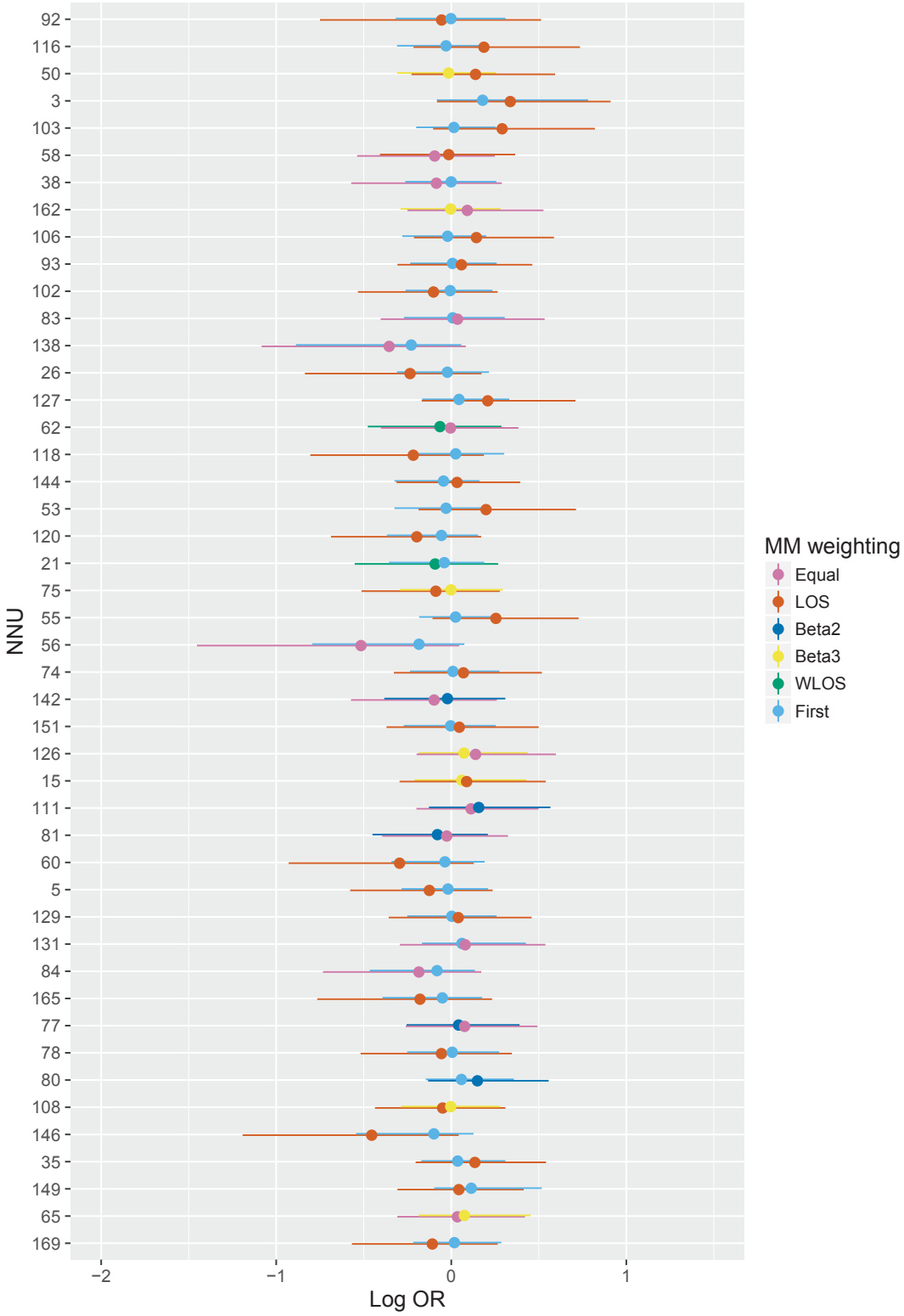


Figure F.3: Caterpillar plots for NICU effects, ordered by proportion transferred

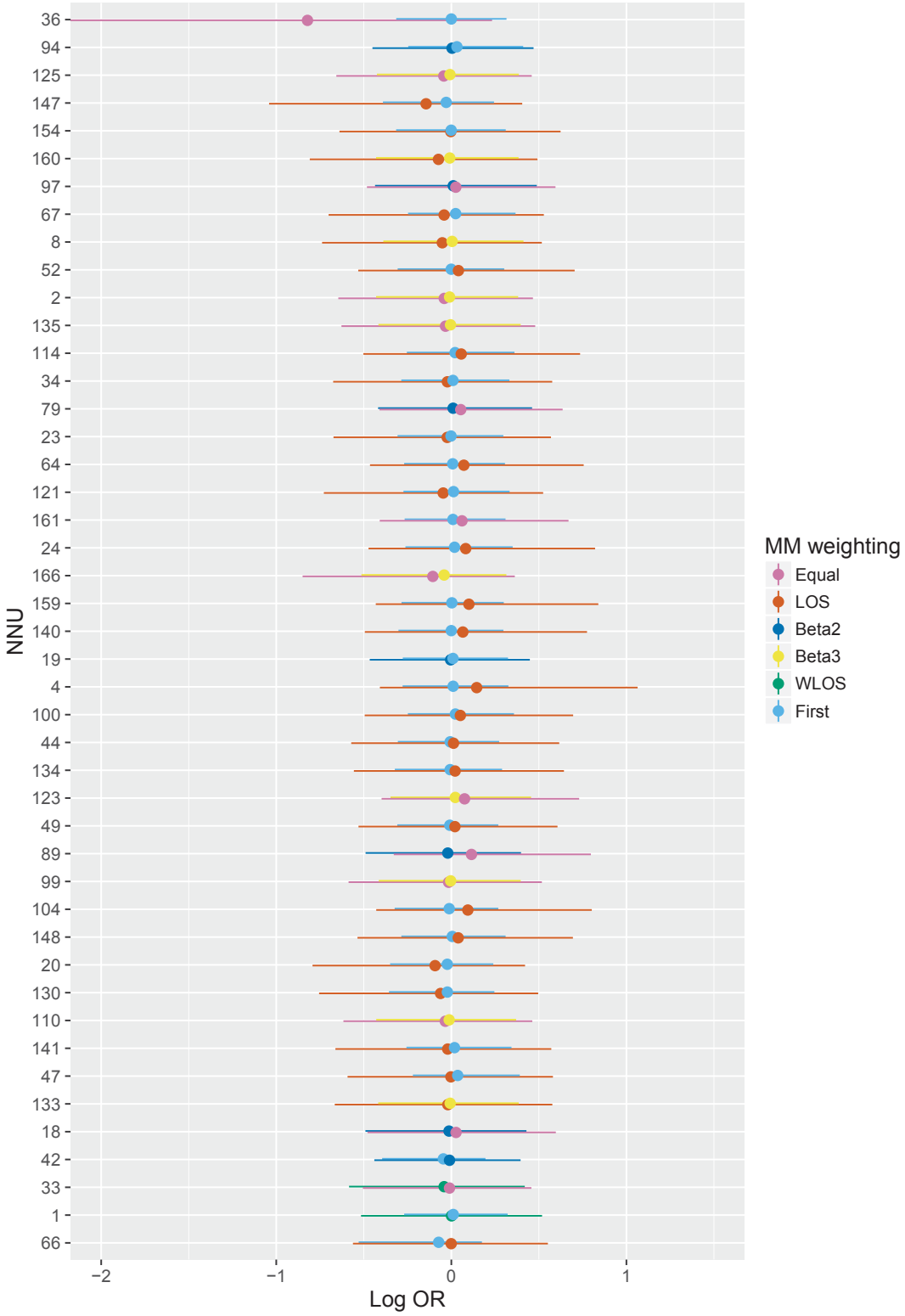


Figure F.4: Caterpillar plots for SCBU effects, ordered by proportion transferred in

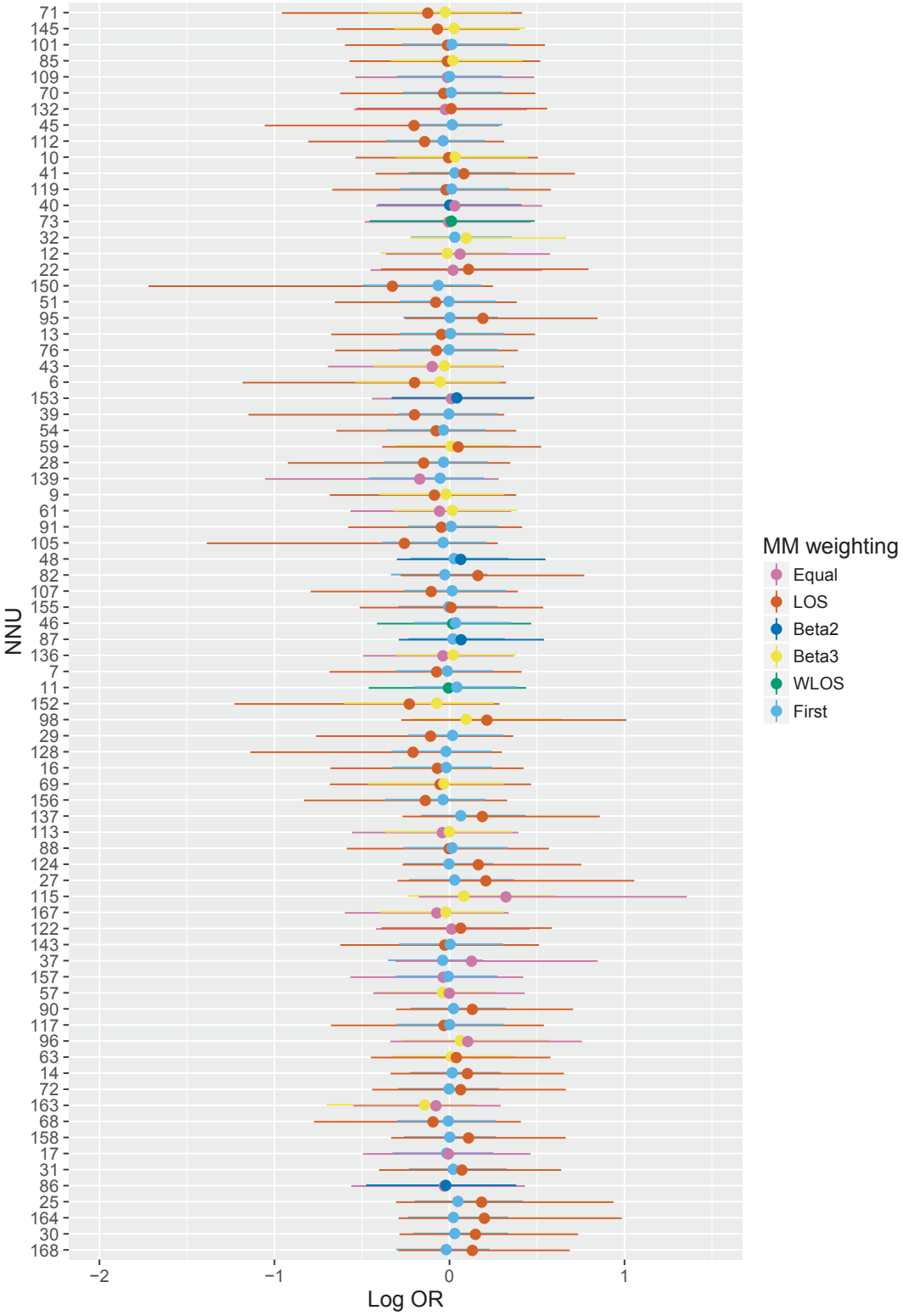


Figure F.5: Caterpillar plots for LNU effects, ordered by proportion transferred in

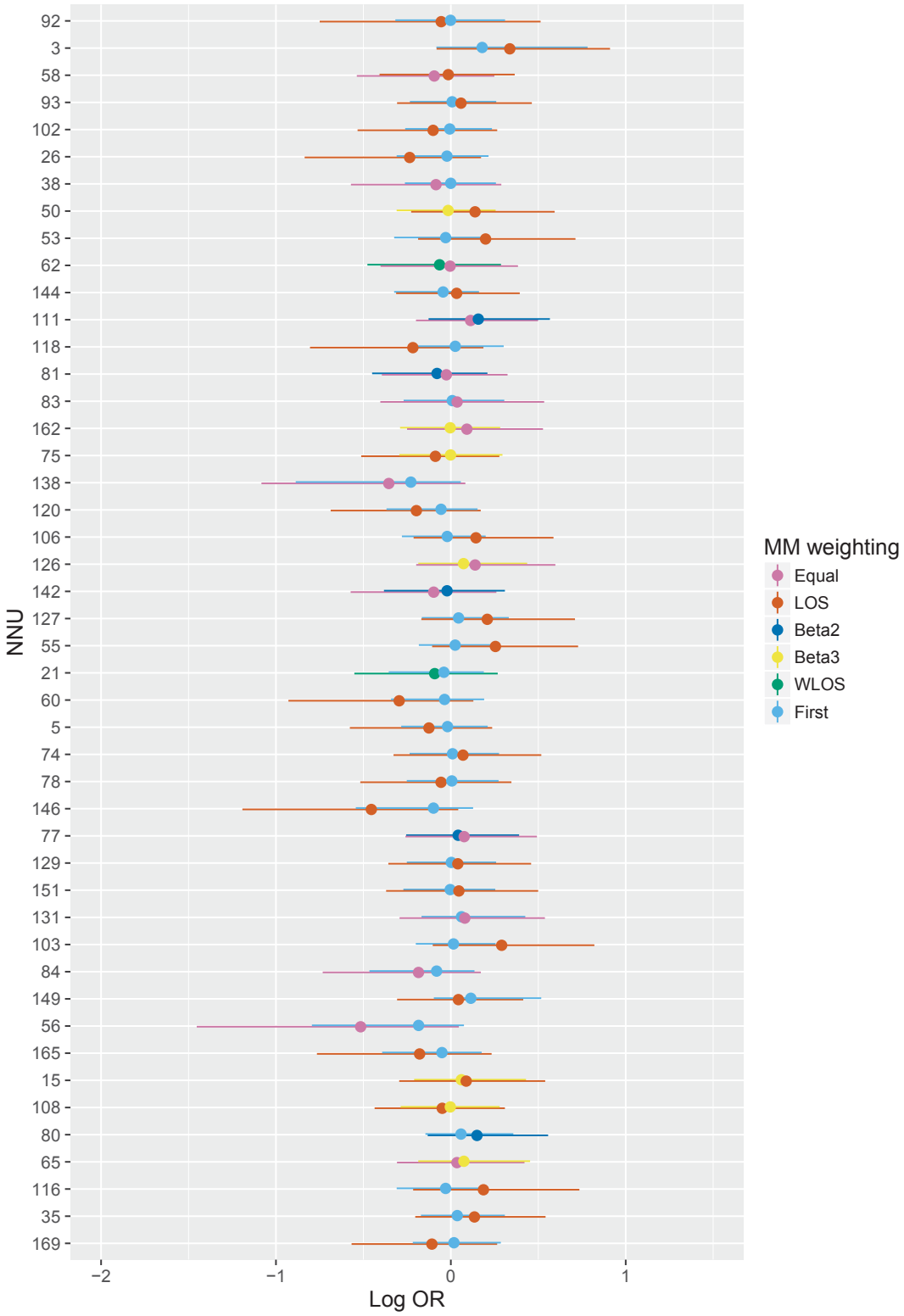


Figure F.6: Caterpillar plots for NICU effects, ordered by proportion transferred in

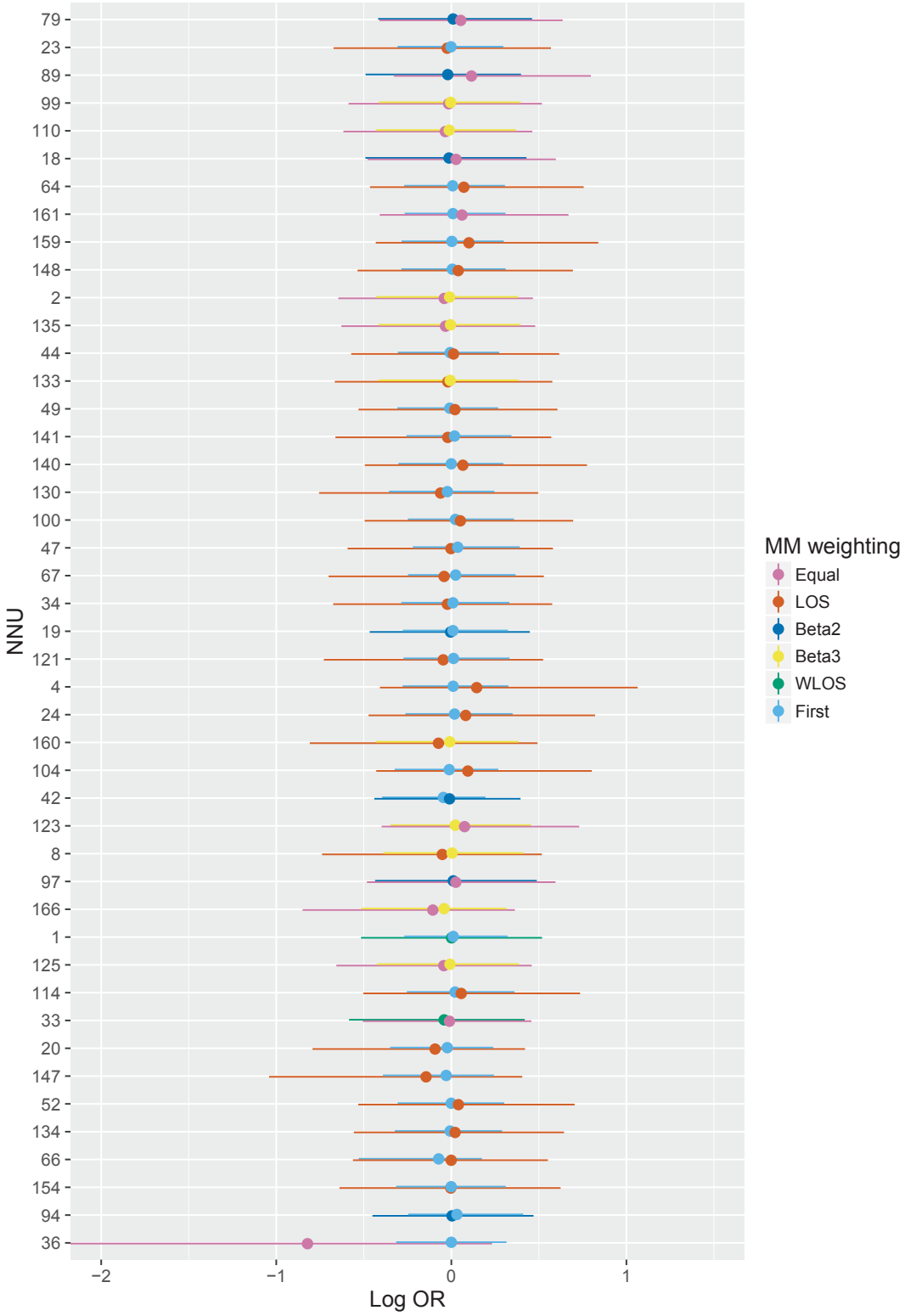


Figure F.7: Caterpillar plots for SCBU effects, ordered by proportion transferred out

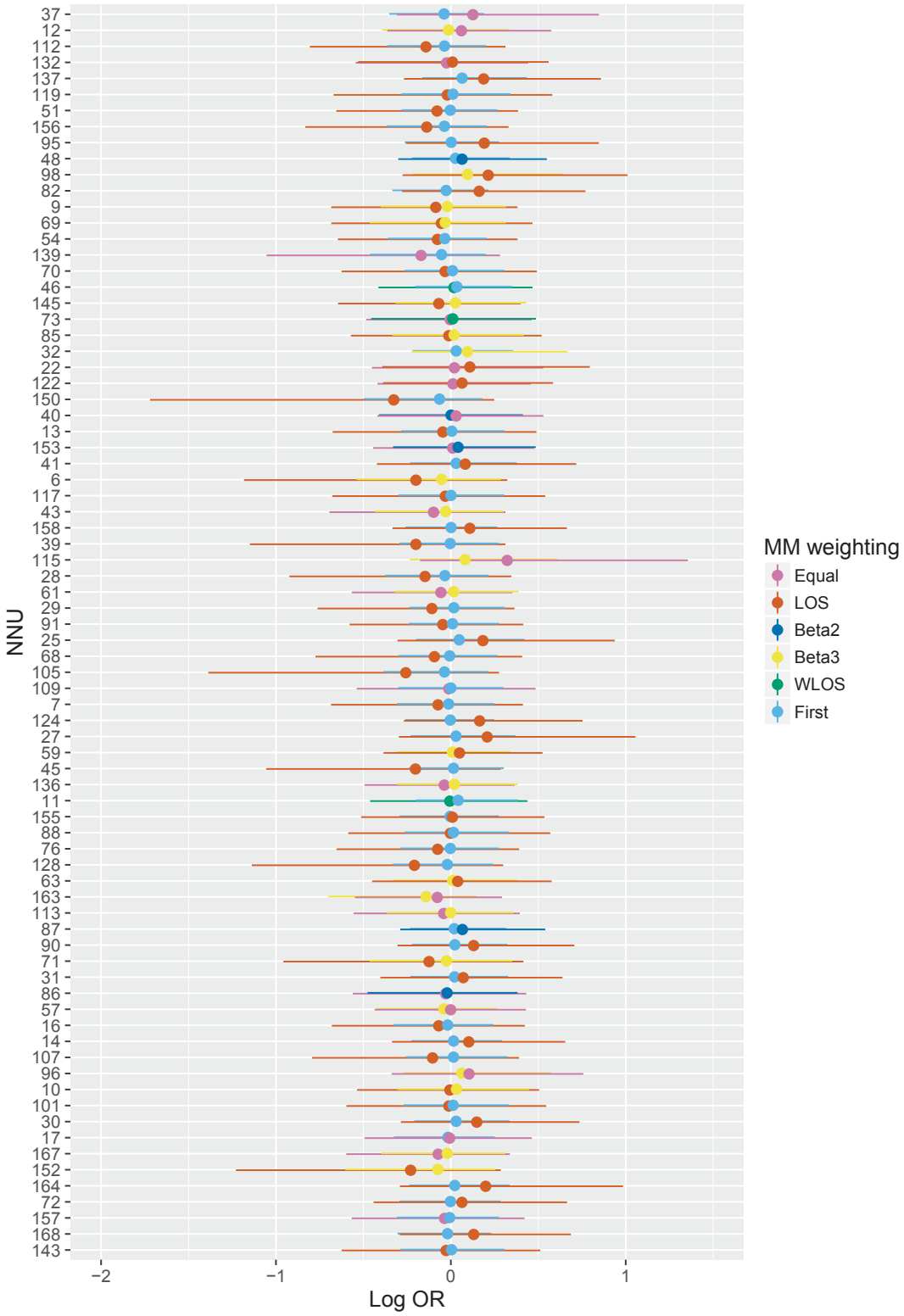


Figure F.8: Caterpillar plots for LNU effects, ordered by proportion transferred out

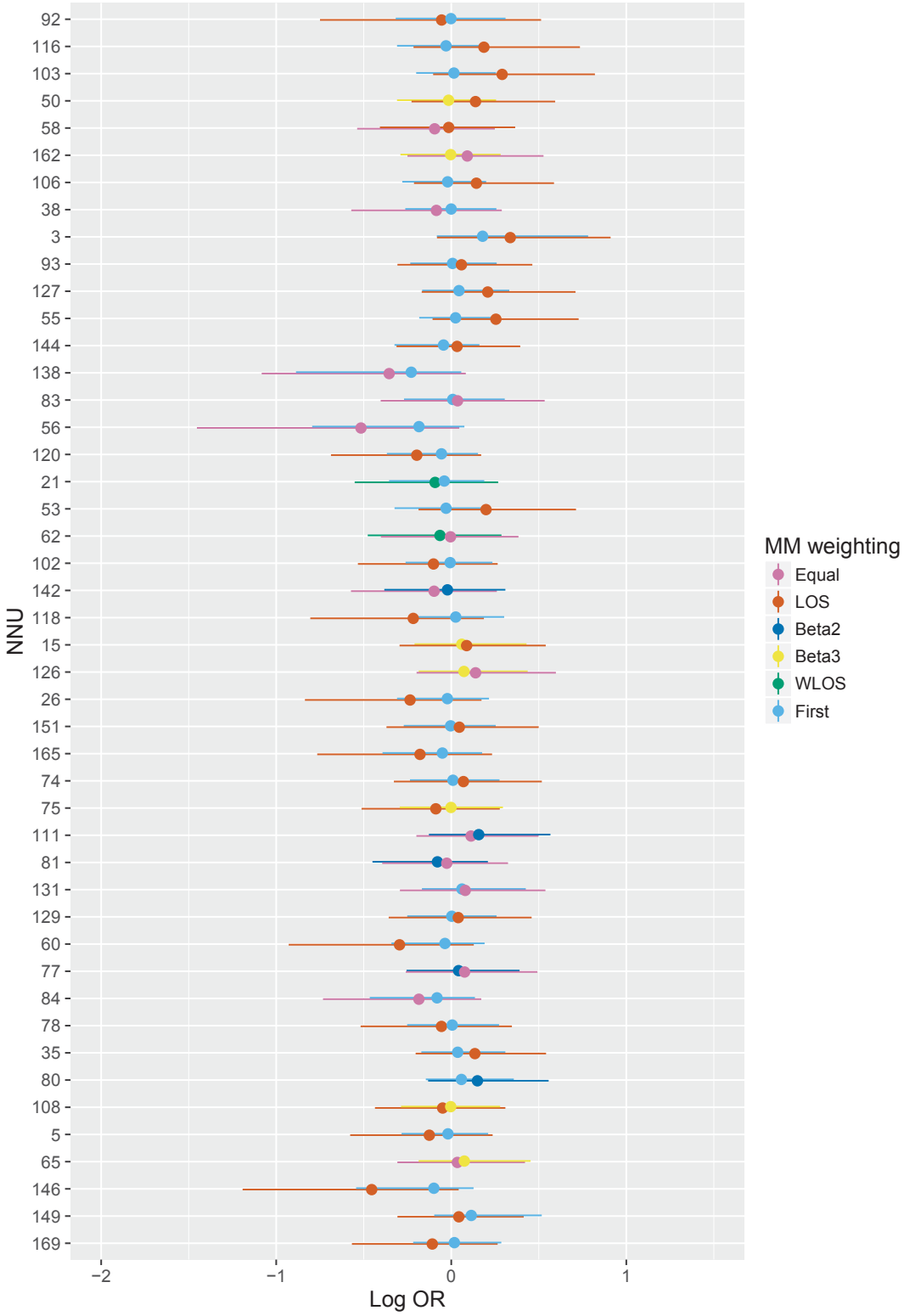


Figure F.9: Caterpillar plots for NICU effects, ordered by proportion transferred out

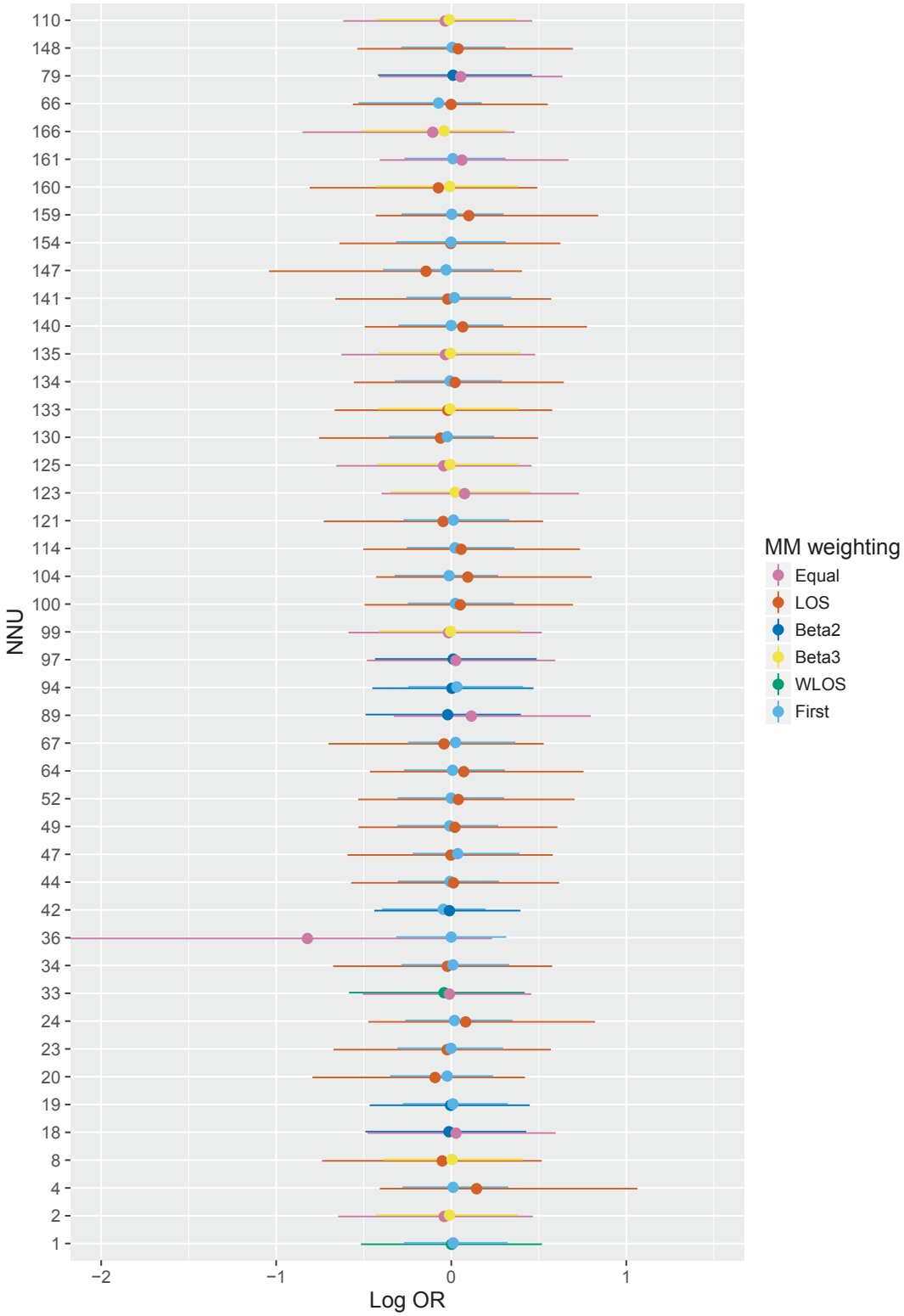


Figure F.10: Caterpillar plots for SCBU effects, ordered by proportion transferred in within 24 hours of birth

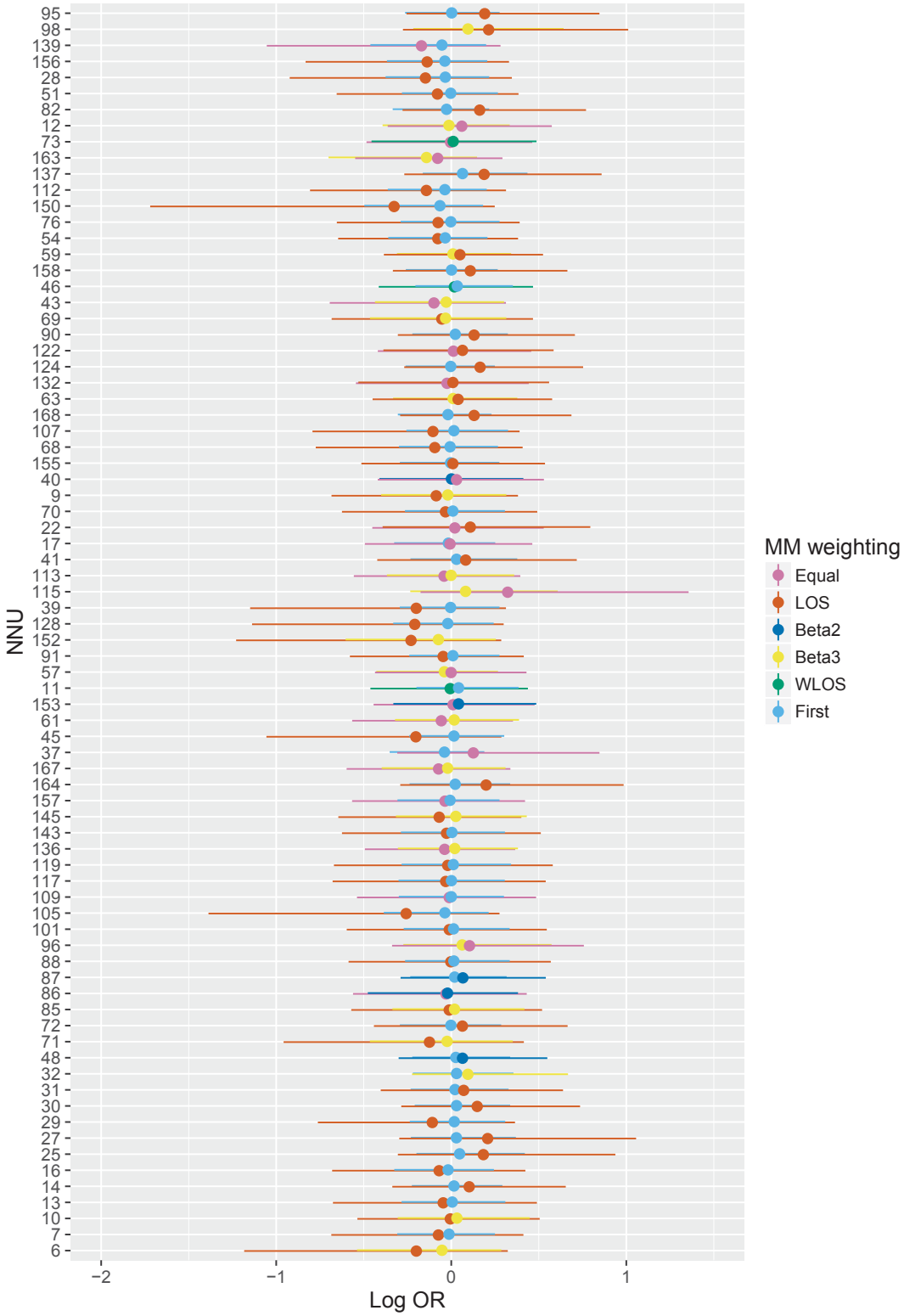


Figure F.11: Caterpillar plots for LNU effects, ordered by proportion transferred in within 24 hours of birth

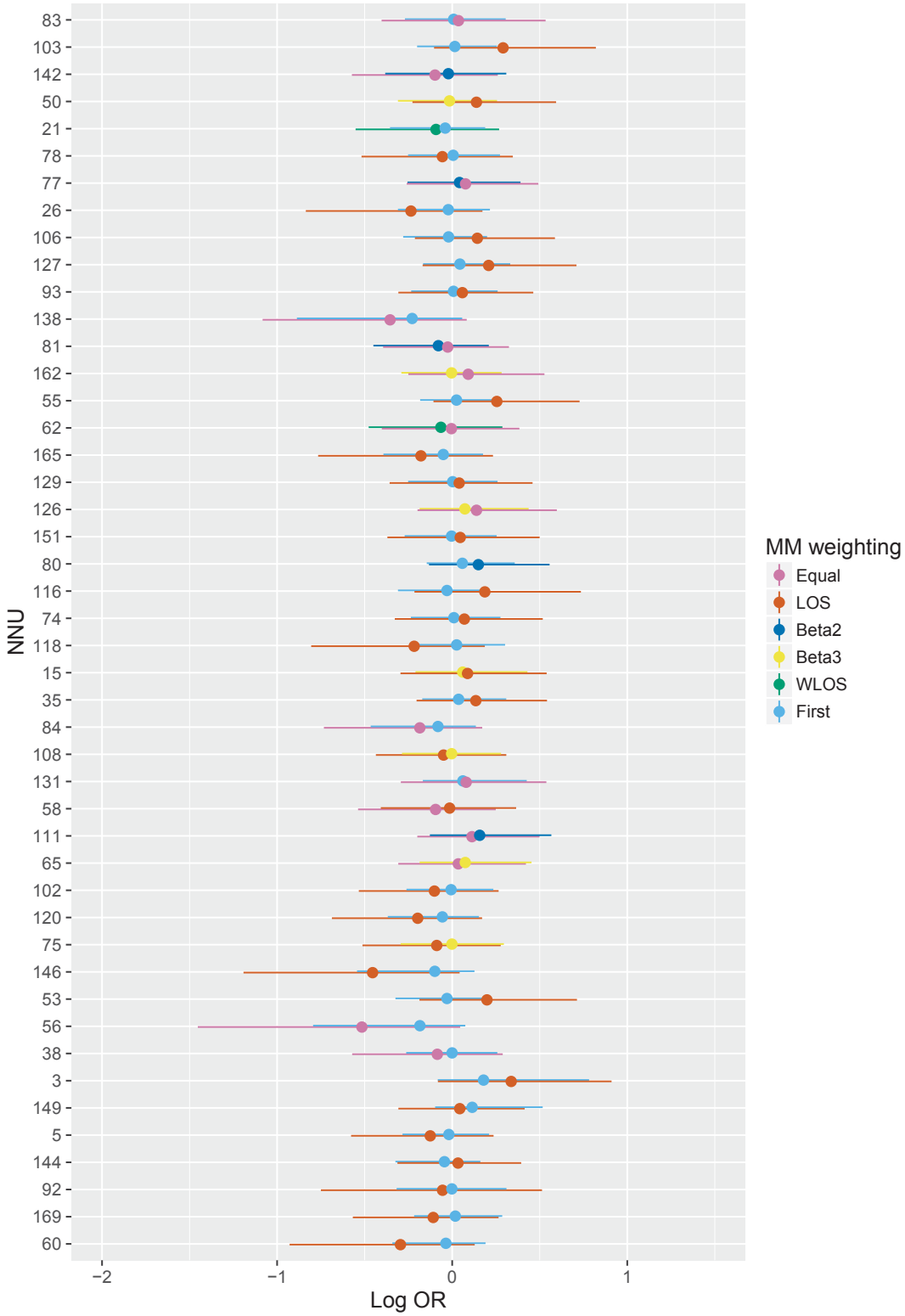


Figure F.12: Caterpillar plots for NICU effects, ordered by proportion transferred in within 24 hours of birth

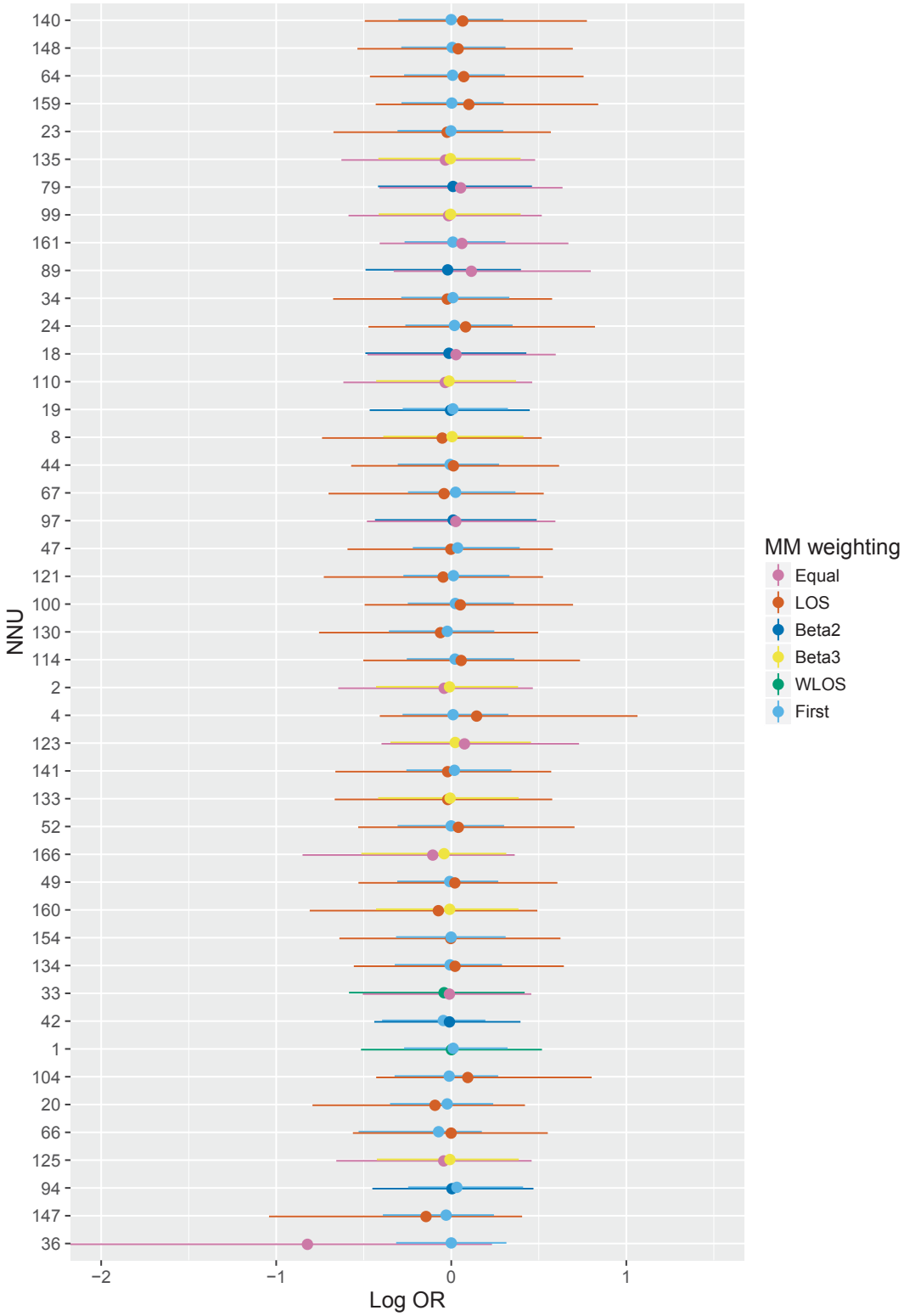


Figure F.13: Caterpillar plots for SCBU effects, ordered by proportion transferred out within 24 hours of birth

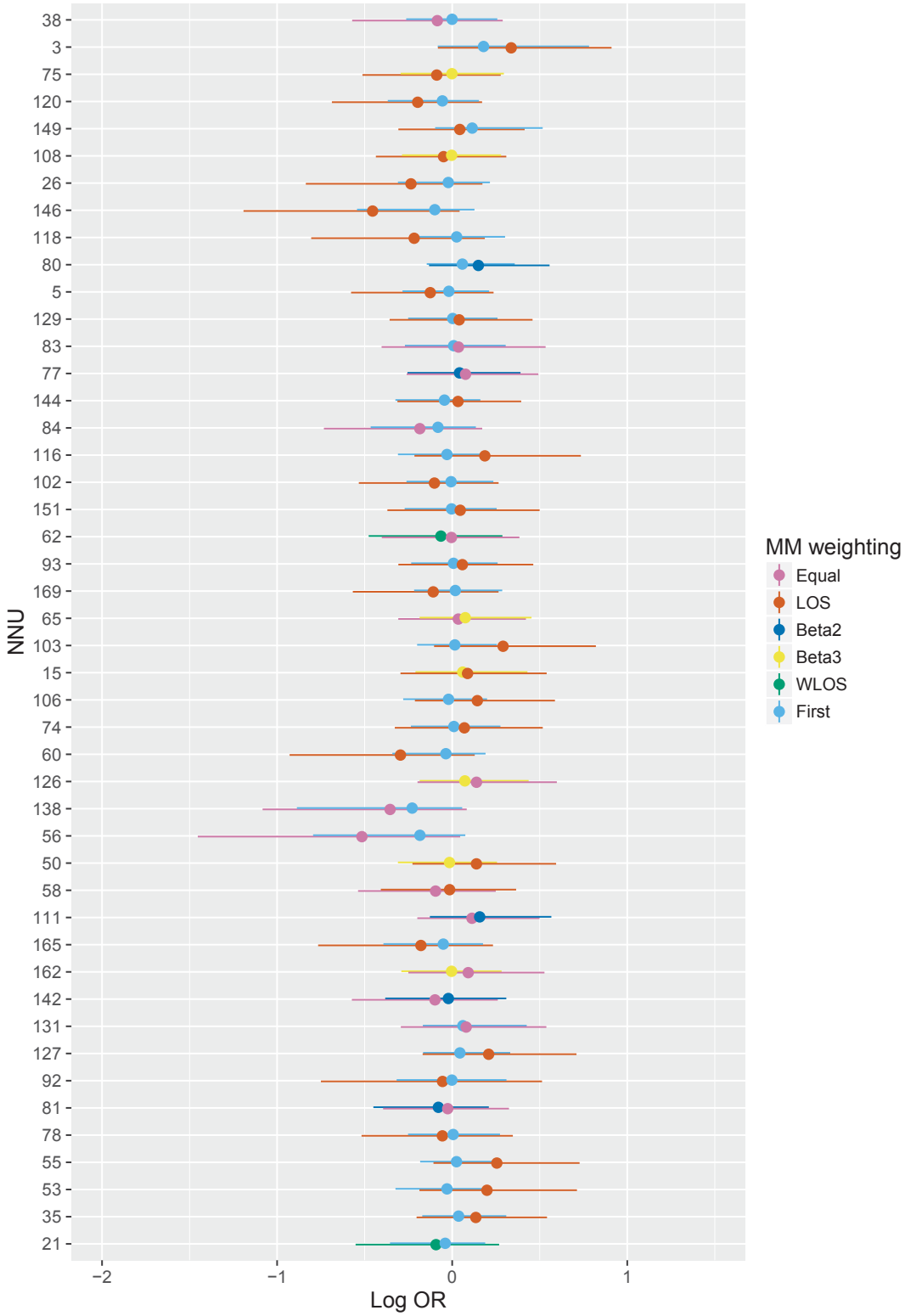


Figure F.14: Caterpillar plots for NICU effects, ordered by proportion transferred out within 24 hours of birth

Appendix G

List of outputs relating to the thesis

Articles related to this work:

S. Santhakumaran, Y. Statnikov, D. Gray, C. Battersby, D. Ashby, and N. Modi on behalf of the Medicines for Neonates Investigator Group. Survival of very preterm infants admitted to neonatal care in England 2008 to 2014: time trends and regional variation *Submitted*

S. Santhakumaran. How to adjust for case-mix when comparing outcomes across healthcare providers. *Arch. Dis. Child. Educ. Pract. Ed.*, 99:70-71, 2014

C. Gale, S. Santhakumaran, S. Nagarajan, Y. Statnikov, and N. Modi. Impact of managed clinical networks on NHS specialist neonatal services in England: population based study *BMJ* 344:e2105, 2012

Other co-authored publications cited in this thesis:

S. I. Watson, W. Arulampalam, S. Petrou, N. Marlow, A. S. Morgan, E. S. Draper, S. Santhakumaran, and N. Modi. The effects of designation and volume of neonatal care on mortality and morbidity outcomes of very preterm infants in England: retrospective population-based cohort study. *BMJ Open* 4(7):e004856, 2014

C. Battersby, S. Santhakumaran, and M. Upton. The impact of a regional care bundle on maternal breast milk use in preterm infants: outcomes of the East of England quality improvement programme. *Arch. Dis. Child. Fetal Neonatal Ed.*, 2014. doi: 10.1136/archdischild-2013-305475.

T. Cole, Y. Statnikov, S. Santhakumaran, and H. Pan. Birth weight and longitudinal growth in infants born below 32 weeks' gestation: a UK population study. *Arch.*

Dis. Child. Fetal Neonatal Ed., 2013. doi: 10.1136/archdischild-2012-303536.

C. Gale, A. Hay, C. Philipp, R. Khan, S. Santhakumaran, and N. Ratnavel. In-utero transfer is too difficult: results from a prospective study. *Early human development*, 88(3):147–50, 2012

Conference abstracts related to this work:

Joint Perinatal Mortality Education Event, Staffordshire Shropshire & Black Country Newborn and Maternity Network, October 2015. Oral presentation: Neonatal Data Analysis Unit 2013 & 2014 Neonatal Mortality Reports

Royal College of Paediatrics and Child Health Annual Conference, June 2013. Oral presentation: Survival of preterm infants admitted to neonatal care in England: a population-based study using NHS electronic clinical records. Awarded a travel grant from Chelsea and Westminster Joint Research Committee

International Society of Clinical Biostatistics, August 2012. Oral presentation: Evaluating mortality rates for neonatal units using multiple membership models

National Neonatal Audit Program Collaborators Meeting, January 2011. Poster and oral presentation: Report on Neonatal Mortality in 2009. Awarded prize for best presentation and a travel grant from Chelsea and Westminster Joint Research Committee