# Clinical prediction modelling to guide decision making for essential neonatal services in Kenyan hospitals




Jalemba Aluvaala

St Catherine's College

University of Oxford

A thesis submitted for the degree of

*Doctor of Philosophy*

Clinical Medicine

Hilary 2018

**Abstract**

*Introduction*

More than 98% of neonatal deaths occur in low and middle income countries(LMICs) and hospitals in these regions are expected to play a key role toward achievement of SDG 3.2 “reduce neonatal mortality to 12 per 1000 live births and below. One approach in efforts to achieve this is to use data for decision making to improve quality of hospital care for neonates.

*Methods*

Using the largest neonatal unit data set in Kenya (11,850 observations): (1) I conducted the first analyses of in-hospital mortality in Kenya using Kaplan-Meier and Competing Risks analysis, (2) developed, validated and explored the potential utility of two novel and simple prognostic models using the Prognosis Research Strategy (PROGRESS) recommendations. Model development was informed by a systematic review guided by the Critical Appraisal and Data Extraction for Systematic Reviews of Prediction Modelling Studies (CHARMS) recommendations.

*Results*

A tenth of the neonates studied died (839/9115, 9.2% [95% C. I 8.7 to 9.8%]). The overall median length of stay was 2 days (range of 0-98 days, IQR 1-5days). Competing risk analysis showed different cumulative incidence of death for the 1.5 to 1.99kg and 2 to 2.49kg weight categories that are often grouped together (1.5 to 2.49). In addition, the competing risk method provides a more clinically meaningful estimate of in-hospital death compared to Kaplan-Meier. Two simple

prognostic models that use treatments (NETS) and signs and symptoms (SENSS) as predictors provide accurate predictions of in-hospital death. Discrimination by c-statistic is 0.89 for each model with calibration slope 0.90 for SENSS and 0.76 for NETS.

*Conclusion*

Prognostic data on in-hospital mortality may inform decisions on treatment, demand and capacity planning as well as balancing of groups in clinical trials. Scale up of such data collection and analyses is required in LMICs to support efforts to improve neonatal survival.

## Declaration

This thesis is entirely my own work except where otherwise indicated.

## Dedication

This work is dedicated first and foremost to my loving wife Ambunya and our two sons Ashivira and Aluvaala. I am eternally grateful for their unwavering support and prayers during my DPhil studies. I also dedicate it to my parents Rose and Meshack Aluvaala for making me be who I am.

**Acknowledgements**

I would like to acknowledge the invaluable mentorship and guidance of my supervisors Prof. Mike English, Prof. Gary S. Collins and Prof. Jay Berkley. I would especially like to thank Prof. Mike English for starting off and continually supporting my research career. Without them this thesis would never have been written.

In addition, I am eternally grateful to my colleagues in the Clinical Information Network from the KEMRI-Wellcome Trust Research Programme, University of Nairobi, Kenya Paediatric Association, Nairobi County and the Ministry of Health, Government of Kenya for supporting the data collection and management system. I would also like to thank the clinical team and administrators of the study hospital and particularly the neonates for providing the data. I am also indebted to the University of Nairobi and especially the Department of Paediatrics and Child Health for affording me time to undertake this work. Last but not least I thank my siblings, parents in-law and the rest of my family for their constant support and encouragement.

## List of figures

Figure 1.1: Definition of the neonatal and perinatal periods. ..... 1
Figure 1.2: Global under-5 case mortality fractions in 2000-15. ..... 3
Figure 1.3: Global population growth by major area. ..... 7
Figure 1.4: Neonatal mortality rate in East Africa 1990-2015. ..... 8
Figure 1.5: Conceptual framework for improving Maternal, Newborn and Child Health (MNCH) across the continuum of care ..... 14
Figure 1.6: Types of clinical prediction models. ..... 27
Figure 1.7: Illustrative search of prediction models indexed on PubMed 1970-2016. ..... 28
Figure 1.8: Phases in the development of prediction models. ..... 29
Figure 2.1: Data management processes ..... 50
Figure 2.2: In-hospital death and discharge alive/referral as competing risks. ..... 55
Figure 3.1: Flow diagram for the selection of studies on neonatal treatment intensity models ..... 84
Figure 4.1: The Economic, Clinical and Humanistic Outcomes model ..... 110
Figure 4.2: Selection of patients for inclusion in the study dataset ..... 118
Figure 4.3: Pattern of admissions to the study neonatal unit ..... 123
Figure 4.4: Distribution of length of stay in the neonatal unit for all the neonates ..... 126
Figure 4.5: Length of stay in the neonatal unit by birthweight for all neonates ..... 127
Figure 4.6: Length of stay by birthweight for neonates who survived to discharge ..... 128
Figure 4.7: Probability of being alive while in hospital over time ..... 130
Figure 4.8: Probability of being alive while in hospital by sex ..... 132
Figure 4.9: Probability of being alive while in hospital by birthweight ..... 133
Figure 4.10: Probability of in-hospital death or discharge alive for all neonates ..... 136
Figure 4.11: Probability of in-hospital death or discharge alive by sex ..... 137
Figure 4.12: Probability of in-hospital death and discharge alive for birthweight less than 1.5kg ..... 139
Figure 4.13: Probability of in-hospital death or discharge alive for birthweight 1.5 to 2.499 kg ..... 140
Figure 4.14: Probability of in-hospital death or discharge alive for birthweight 1.5 to 1.999 and 2 to 2.499kg ..... 142
Figure 4.15: Probabilty of in-hospital death or discharge alive for birthweight 2.5 to 4k and greater than 4 kg ..... 143
Figure 4.16: Comparison of cumulative incidence of in-hospital mortality by Kaplan-Meier and Competing Risks methods ..... 150
Figure 5.1: Patient flow chart for NETS model derivation ..... 163
Figure 5.2: Patient flow chart for NETS model external validation ..... 170
Figure 5.3: Calibration curve for NETS in the derivation data set ..... 178
Figure 5.4: Calibration curve for NETS in the external validation data set ..... 185
Figure 5.5: Calibration curve for NETS model after external validation and updating ..... 187
Figure 6.1: Calibration plot for SENSS in the derivation data set ..... 213
Figure 6.2: Calibration curve for SENSS in the external validation data set ..... 216
Figure 6.3: Calibration curve for the final SENSS model update of the intercept ..... 219
Figure 6.4: Performance of NETS and SENSS at external validation in the same population ..... 226
Figure 7.1: Comparison of the standardised net benefit treated for NETS and SENSS models in predicting in-hospital mortality ..... 243
Figure 7.2: Clinical impact of cut-offs across the range of predictions for NETS (left) and SENSS (right) ..... 247

## List of tables

Table 1.1: Global variation in estimated neonatal mortality rate for 2015 including the country specific estimate for Kenya. ..... 6
Table 1.2: Ten countries with the highest neonatal mortality rate in sub-Saharan Africa in 2015. ..... 6
Table 1.3: Progress in under-5 and neonatal mortality in Kenya 2000-15 ..... 9
Table 1.4: Neonatal mortality rate by region in Kenya for 2009-2014 ..... 10
Table 1.5: Illustration of how information can guide decision making processes in hospital care for neonates in Kenya ..... 17
Table 1.6: Birth weight categories ..... 18
Table 1.7: Gestational age categories ..... 18
Table 1.8: Neonatal inpatient case fatality in Kenyan hospitals ..... 24
Table 1.9: Potential application of prediction models in neonatal medicine in Kenya ..... 31
Table 2.1 : Research objectives under the PROGRESS prognostic research themes ..... 35
Table 2.2: Key items in the design of systematic reviews of prediction modelling studies. ..... 39
Table 2.3: Relevant items for data extraction and critical appraisal. ..... 40
Table 2.4: Ranking the importance of outcomes using GRADE. ..... 43
Table 2.5: Evaluation of certainty in the evidence across studies using GRADE. ..... 44
Table 2.6: Application of GRADE to assess certainty of evidence in prediction modelling studies ..... 45
Table 2.7: Rating of certainty of evidence in predictive model predictive performance ..... 46
Table 2.8: Staff on duty in the newborn unit on a typical shift in the study hospital ..... 48
Table 2.9: Description of the study data sets ..... 52
Table 2.10: Summary of methods used in prediction model development and validation ..... 56
Table 3.1: Examples of “medical” models used to grade severity of illness in neonates 72
Table 3.2: Summary of prognostic models for predicting in-hospital neonatal mortality ..... 75
Table 3.3: Eligibility criteria for inclusion in the systematic review ..... 80
Table 3.4: Characteristics of studies included in the review ..... 86
Table 3.5: Limitations in individual studies with respect to study design and data collection ..... 88
Table 3.6: Limitations in individual studies with respect to statistical analysis ..... 90
Table 3.7: Certainty in the evidence of predictive performance across all studies ..... 95
Table 3.8: Summary of results from individual studies ..... 98
Table 4.1: Case fatality ratio in neonatal units of African hospitals ..... 113
Table 4.2: Clinical characteristics of neonates included in the study data set ..... 124
Table 5.1:Therapeutic interventions used for in-patient, post-natal essential newborn care ..... 158
Table 5.2: Essential treatments prescribed at admission that were included as predictors ..... 165
Table 5.3: Additional predictors included in the treatment intensity model ..... 166
Table 5.4: Clinical characteristics of neonates included in the model development data set ..... 173
Table 5.5: Distribution of predictors by in-hospital mortality in the NETS derivation dataset ..... 174
Table 5.6: Coefficients from the logistic regression model for the NETS score ..... 176
Table 5.7: Overall measures of NETS model apparent performance ..... 177
Table 5.8: Calibration of the NETS model after correction for optimism ..... 181
Table 5.9: Discrimination of the NETS model after correction for optimism ..... 181

Table 5.10: Shrinkage of NETS model coefficients to correct for optimism ..... 182
Table 5.11: Characteristics of patients included in NETS derivation and external validation ..... 183
Table 5.12: Distribution of predictors by in-hospital mortality ..... 184
Table 5.13: Final NETS model after updating ..... 186
Table 5.14: Summary of the predictive accuracy of NETS ..... 188
Table 6.1: Clinical predictors of severe illness in young infants (0-60 days) ..... 198
Table 6.2: Symptoms and signs of severe illness considered as candidate predictors 202
Table 6.3: Distribution of predictors and missingness in the SENSS derivation dataset ..... 203
Table 6.4: Distribution of predictors and missingness in the external validation dataset ..... 207
Table 6.5: Distribution of predictors in the original and imputed data sets ..... 209
Table 6.6: Distribution of predictors by in-hospital mortality in the imputed datasets .... 210
Table 6.7: Coefficients from the logistic regression model for the NETS score ..... 212
Table 6.8: Overall measures of NETS model apparent performance ..... 212
Table 6.9: Performance of the SENSS model after internal validation ..... 214
Table 6.10: Discrimination of the SENSS model after correction for optimism ..... 215
Table 6.11: Shrinkage of the SENSS model coefficients to correct for optimism ..... 215
Table 6.12: Coefficients of the logistic regression model for the updated SENSS score ..... 217
Table 6.13: Summary of the predictive accuracy of SENSS ..... 221
Table 6.14: Distribution of predictors by in-hospital mortality for NETS and SENSS .... 229
Table 7.1: Factors affecting clinical uptake of prediction models ..... 233
Table 7.2: Computation of simplified points from the model coefficients ..... 241
Table 7.3: Estimated annual referrals at different cut-offs for NETS and SENSS in three Nairobi County hospitals ..... 249
Table 7.4: Further simplified points system for the NETS model ..... 250
Table 7.5: Comparison of predicted risk by the original and simplified points for NETS 251
Table 7.6: NETS final simplified score chart ..... 252
Table 7.7: NETS patient score sheet ..... 252
Table 7.8: SENSS final score chart ..... 254
Table 7.9: SENSS patient score sheet ..... 254
Table 7.10: Predicted risk of in-hospital death using the NETS and SENSS score charts for six cases randomly drawn from the validation dataset ..... 256
Table 8.1: Case fatality by birthweight in the study hospital versus two high income setting neonatal networks ..... 269
Table 8.2: Examples of neonatal prognostic factors in the RMNCH continuum of care 289

**List of text boxes**

Box 1.1: An example of a prediction model formula ..... 29
Box 2.1: The Grading of recommendations assessment, development, and evaluation (GRADE) process ..... 42
Box 2.2: Estimation of the linear predictor of death in logistic regression ..... 58
Box 2.3: Calculation of predicted probability of death from odds of death ..... 58
Box 2.4: Calculation of optimism-corrected performance for prediction models. ..... 64
Box 2.5: Calculation of net benefit across the range of threshold probabilities ..... 67
Box 2.6: Illustration of calculation of net benefit at a given cut-off ..... 68
Box 5.1: Logistic regression model for the NETS score ..... 175
Box 5.2: NETS derivation regression formula for the linear predictor of in-hospital mortality ..... 176
Box 5.3: Final logistic regression model for NETS after updating the intercept ..... 193
Box 6.1: Logistic regression model for the SENSS score ..... 211
Box 6.2: Final logistic regression model for SENSS after updating the intercept ..... 218
Box 7.1: Development of a points system for a multivariable logistic regression model 236
Box 7.2: Conversion of model points to predicted probability of in-hospital death ..... 241
Box 7.3: Calculation of net benefit untreated across the range of threshold probabilities ..... 261

## List of abbreviations

| | |
|---|---|
| **AAP** | American Academy of Pediatrics |
| **AJOL** | African Journals Online |
| **ANZNN** | Australian and New Zealand Neonatal Network |
| **AUROC** | Area Under the Receiver Operating Characteristic Curve |
| **BCPAP** | Bubble Continuous Positive Airway Pressure |
| **CFR** | Case Fatality Ratio |
| **CHARMS** | Checklist for Critical Appraisal and Data Extraction for Systematic Reviews of Prediction Modelling Studies |
| **CNN** | Canadian Neonatal Network |
| **COIN** | Core Outcomes in Neonatology |
| **COMET** | Core Outcome Measures in Effectiveness Trials |
| **CRIB** | Clinical Risk Index for Babies |
| **DHIS2** | District Health Information System Version 2 |
| **EAC** | East African Community |
| **ELBW** | Extremely Low Birth Weight (<1kg) |
| **ENAP** | Every Newborn Action Plan |
| **EPV** | Events Per Variable |
| **GRADE** | Grading of Recommendations Assessment, Development, And Evaluation |
| **H-L** | Hosmer-Lemeshow |
| **ICD-PM** | International Classification of Diseases – Perinatal Mortality |
| **IUGR** | Intra-Uterine Growth Restriction |
| **KDHS** | Kenya Demographic and Health Survey |
| **KEMRI** | Kenya Medical Research Institute |
| **LICs** | Low Income Countries |
| **LMICs** | Low and Middle Income Countries |
| **MAR** | Missing at Random |
| **MCAR** | Missing Completely at Random |
| **MDG** | Millennium Development Goal |
| **MeSH** | Medical Subject Headlines |
| **MNAR** | Missing Not at Random |
| **NAR** | Neonatal Admission Record |
| **NB** | Net benefit |
| **NBRS** | Nursery Neurobiologic Risk Score |
| **NEOCOSUR** | Neonatal Del Cono Sur Network |
| **NEOMOD** | Neonatal Multiple Organ Dysfunction |
| **NETS** | Neonatal Essential Treatment Score |
| **NICE** | National Institute of Health and Care Excellence |
| **NICHHD** | National Institute of Child Health and Human Development |
| **NICU** | Neonatal Intensive Care Unit |
| **NMPI** | Neonatal Mortality Prognosis Index |
| **NMR** | Neonatal Mortality Rate |
| **NQI-2** | Neonatal Quality Indicator |
| **NTISS** | Neonatal Treatment Intensity Scoring System |
| **OxTREC** | Oxford Tropical Research Ethics Committee |
| **PAQC** | Paediatric Admission Quality of Care |
| **PICOT** | Population Issue Context Outcome Time |

| | |
|---|---|
| **PROBAST** | Prediction Study Risk of Bias Assessment Tool |
| **PROGRESS** | Prognosis Research Strategy |
| **PROSPERO** | Prospective Register of Systematic Reviews |
| **REDCap** | Research Electronic Data Capture |
| **RoB** | Risk of Bias |
| **SAW** | Sex Age Weight |
| **sNB** | Standardised Net Benefit |
| **SDG** | Sustainable Development Goals |
| **SENSS** | Score for Essential Neonatal Symptoms and Signs |
| **SGA** | Small for Gestational Age |
| **SNAP** | Score for Neonatal Acute Physiology |
| **SNAP-PE** | Score for Neonatal Acute Physiology-Perinatal Extension |
| **STROBE-NI** | Strengthening The Reporting of Observational Studies In Epidemiology for Newborn Infection |
| **THIN** | The Health Improvement Network |
| **TISS** | Therapeutic Intervention Scoring System |
| **TRIPOD** | Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis Or Diagnosis |
| **TRIPS** | Transport Risk Index of Physiologic Stability |
| **U5M** | Under-Five Mortality Rate |
| **VLBW** | Very Low Birth Weight |
| **VON-RA** | Vermont Oxford Network Risk Adjustment |
| **WHO** | World Health Organisation |

## Contents

**Chapter 1 : Introduction ..... 1**

**1.1 Neonatal survival ..... 1**

**1.2 Interventions to improve neonatal survival and the role of hospital care ..... 11**

**1.3 Use of information in decision making to improve quality of hospital care ..... 16**

**1.4. Role of prognostic data in decision making to improve quality of health care ..... 21**

**1.5 Prognosis after neonatal admission in Kenyan hospitals ..... 23**

**1.6 Role of clinical prediction models in supporting decision making ..... 27**

**1.7 Research objectives ..... 33**

**1.8 Justification ..... 34**

**Chapter 2 : Methods ..... 35**

**2.1 Introduction ..... 35**

**2.2 Systematic review of prediction modelling studies ..... 37**

**2.3 Study site and data collection procedures ..... 47**

**2.4 Description of the study data sets ..... 52**

**2.5 Survival analysis ..... 53**

**2.6 Prognostic modelling ..... 56**

**2.7 Clinical utility of prediction models ..... 66**

**2.8 Reporting of prediction modelling studies ..... 69**

**2.9 Ethical considerations ..... 69**

**Chapter 3 : Systematic review of neonatal treatment intensity scores and their potential application for low resource settings ..... 70**

**3.1 Introduction ..... 70**

3.1.1 Role of clinical prediction models in decision making for essential neonatal care ..... 70

3.1.2 Scoping review of prediction models in neonatal medicine ..... 71

3.1.3 Objectives of the systematic review ..... 78

**3.2 Methods ..... 79**

3.2.1 Overview of methods ..... 79

3.2.2 Eligibility criteria ..... 80

3.2.3 Information sources ..... 80

3.2.4 Search strategy ..... 81

3.2.5 Study Selection ..... 82
3.2.6 Data Collection Process ..... 82
3.2.7 Data Items ..... 82
3.2.8 Critical appraisal ..... 83
3.2.9 Synthesis of results ..... 83

**3.3 Results ..... 84**
3.3.1 Study Selection ..... 84
3.3.2 Study characteristics ..... 85
3.3.3 Critical appraisal of individual studies ..... 86
3.3.4 Critical appraisal of certainty of the evidence in score performance for outcomes across all studies ..... 93
3.3.5 Synthesis of results of estimated performance ..... 96

**3.4 Discussion ..... 103**
3.4.1 Performance of neonatal treatment intensity scores ..... 103
3.4.2 Generalisability to essential neonatal care in low resource settings ..... 106

**Chapter 4 : Case fatality ratio, length of stay and time to in-hospital death or discharge ..... 109**

**4.1 Introduction ..... 109**
4.1.1 Background ..... 109
4.1.2 Case fatality and length of stay in neonatal units in Kenya and other African countries ..... 112
4.1.3 Rationale ..... 115
4.1.4 Objective ..... 116

**4.2 Methods ..... 117**
4.2.1 Study design ..... 117
4.2.2 Setting ..... 117
4.2.3 Participants ..... 117
4.2.4 Variables ..... 118
4.2.5 Data sources and measurement ..... 119
4.2.6 Bias ..... 119
4.2.7 Study size ..... 120
4.2.8 Quantitative variables ..... 120
4.2.9 Statistical methods ..... 120

**4.3 Results ..... 123**
4.3.1 Patterns of admissions during the study period ..... 123
4.3.2 Characteristics of the study population and overall case fatality rate ..... 124
4.3.3 Length of stay in the neonatal unit ..... 126
4.3.4 Kaplan-Meier survival analysis of time to in-hospital death ..... 130
4.3.5 Competing risk survival analysis ..... 135

**4.4 Discussion ..... 144**
4.4.1 Summary of findings ..... 144
4.4.2 Case fatality rate ..... 145
4.4.3 Length of stay ..... 147
4.4.4 Time to in-hospital death ..... 149
4.4.5 Categorisation of birthweight for average and individual prognosis ..... 153
4.4.6 Limitations ..... 153
4.4.7 Conclusion ..... 155

**Chapter 5 : Development and validation of the Neonatal Essential Treatment Score (NETS) ..... 156**

**5.1: Introduction ..... 156**
5.1.1 Background ..... 156
5.1.2 Essential neonatal treatments as predictors of in-hospital mortality ..... 157
5.1.3 Rationale ..... 159
5.1.4 Objective ..... 160

**5.2 Methods ..... 161**
5.2.1 Derivation ..... 161
5.2.2 Internal validation ..... 169
5.2.3 External Validation ..... 169

**5.3 Results ..... 173**
5.3.1 Clinical characteristics of the study patients ..... 173
5.3.2 Model derivation ..... 174
5.3.3 Internal validation ..... 180
5.3.4. External validation ..... 182

**5.4 Discussion ..... 188**
5.4.1 Summary of findings ..... 188
5.4.2 Model specification and estimation in derivation ..... 189
5.4.3 Predictive performance at derivation ..... 189
5.4.4 Predictive performance at internal validation ..... 190
5.4.5 Performance at external validation ..... 191
5.4.6 Model updating ..... 193
5.4.7 Final model and clinical utility ..... 193
5.4.8 Limitations ..... 194
5.4.9 Implications ..... 195

**Chapter 6 : Development and validation of the Score of Essential Neonatal Symptoms and Signs (SENSS) ..... 196**

**6.1 Introduction ..... 196**
6.1.1 Background ..... 196
6.1.2 Signs and symptoms of severe illness in neonates as prognostic factors ..... 197
6.1.3 Rationale ..... 199
6.1.4 Objective ..... 200

**6.2 Methods ..... 201**
6.2.1 Model derivation ..... 201
6.2.2 Internal validation ..... 206
6.2.3 External Validation ..... 206

**6.3 Results ..... 209**
6.3.1 Description of the multiply imputed data sets ..... 209
6.3.2 Model derivation ..... 211
6.3.3 Internal validation ..... 214
6.3.4 External validation ..... 216

**6.4 Discussion ..... 220**
6.4.1 Summary of findings ..... 220

6.4.2 Predictor selection ..... 222
6.4.3 Comparison of performance between NETS and SENSS at external validation ..... 225
6.4.4 Potential reasons for difference in calibration between NETS and SENSS ..... 226
6.4.5 Justification for external validation ..... 229
6.4.6 Limitations ..... 230
6.4.7 Implications ..... 230

**Chapter 7 : Evaluation of the utility of the NETS and SENSS models in clinical decision making ..... 232**

**7.1 Introduction ..... 232**
7.1.1 Background ..... 232
7.1.2 Factors affecting the adoption of clinical prediction models into routine practice ..... 232
7.1.3 The role of decision curve analysis and presentation format in the clinical utility of prognostic models ..... 234
7.1.4 Rationale ..... 236
7.1.5 Objective ..... 237

**7.2 Methods ..... 238**
7.2.1 Decision curve analysis ..... 238
7.2.2 Simplified points based system ..... 240

**7.3 Results ..... 243**
7.3.1 Decision curve analysis evaluating clinical utility of the NETS and SENSS models ..... 243
7.3.2 NETS and SENSS presentation as score charts ..... 250

**7.4 Discussion ..... 258**
7.4.1 Summary of findings ..... 258
7.4.2 Clinical utility of prediction models ..... 259
7.4.3 Clinical presentation of prediction models ..... 265
7.4.4 Limitations ..... 265
7.4.5 Implications ..... 266

**Chapter 8 : Discussion ..... 267**

**8.1 Summary of findings ..... 267**
8.1.1 Fundamental prognosis ..... 267
8.1.2. Prognostic factors and prognostic modelling ..... 270

**8.2 Implications for neonatal routine data and hospital based prognosis research ..... 272**
8.2.1 The potential of neonatal routine data in LMICs ..... 272
8.2.2 Standardising neonatal outcomes in research ..... 274
8.2.3 Impact on neonatal prognostic research ..... 275

**8.3 Prognostic data for decision making to improve neonatal survival in Kenya ..... 276**
8.3.1 Clinical decision making for individual patients ..... 276
8.3.2 Describing and comparing cohorts of patients admitted to neonatal units ..... 281
8.3.3 Service delivery planning and organisation ..... 283

**8.4 Strengths and limitations ..... 285**

**8.5 Recommendations and future directions ..... 287**
8.5.1 Fundamental prognosis ..... 287

8.5.2 Prognostic factor research ..... 288
8.5.3 Prognostic modelling ..... 289

**8.6 Conclusion ..... 292**

**References ..... 293**

**Appendix ..... 312**

## Chapter 1 : Introduction

### 1.1 Neonatal survival

*1.1.1 Definition of the neonatal and perinatal periods*

The biological vulnerability in the time periods surrounding pregnancy and child birth (Figure 1.1) results in high risk of morbidity and mortality to the mother, foetus and newborn child. Birth occurs after varying lengths of pregnancy: preterm birth after less than 37 completed weeks of pregnancy, term 37-42 weeks, post-dates after 42 weeks. The neonatal or newborn period refers to the first 27 completed days of life after birth regardless of the duration of pregnancy (Figure 1.1). The perinatal period spans from 22 completed weeks of pregnancy to 6 days after birth and combines foetal and neonatal events.

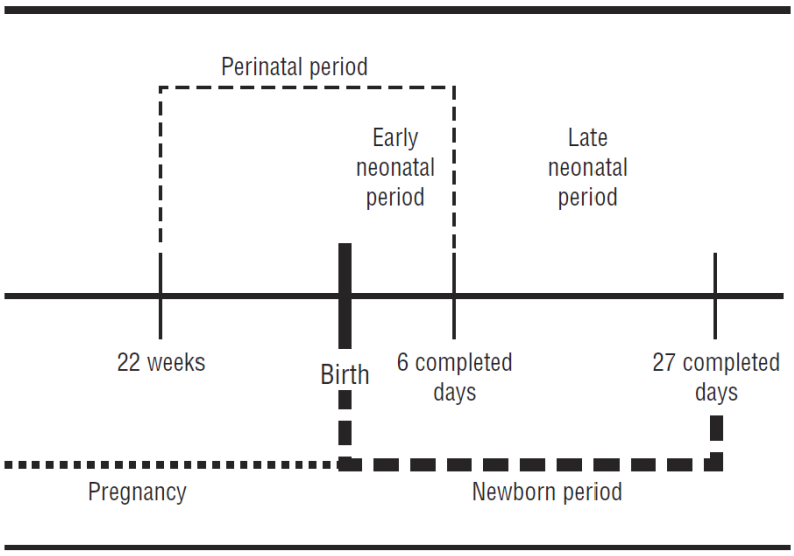



***Figure 1.1:*** *Definition of the neonatal and perinatal periods. Reproduced from Indira N, Mandy R, Dilberth C et al. The Components of Essential Newborn Care. 2004*

Globally, the highest risk of death in childhood (age 0-18 years) is observed in the neonatal period and in particular, the first 6 days – the early neonatal period.(1, 2) Consequently, the subsequent sections of this thesis focus predominantly on

mortality in this high risk period specifically for those admitted to newborn units. This population tends to have become sick at the time of birth or in the first 24 to 48 hours after birth

*1.1.2 Global estimates of neonatal mortality*

The Millennium Development Goal (MDG) 4 set a target of two thirds reduction in the under-five mortality rate (U5M) between 1990-2015.(3) By 2015, the global U5M was reduced by half from 90 to 43 deaths per 1,000 live births.(3) However, only 16 countries in sub-Saharan Africa achieved the target annual rate of reduction of 4.4% per year in the period 2000-15.(2) The MDG4 did not include a specific neonatal target, a group for whom slower improvement in survival is seen; 42% reduction in the global neonatal mortality rate (NMR) from 33 to 19 deaths per 1,000 live births in 1990 and 2015 respectively. This represents an annual reduction of 3.1% in contrast to 4·0% and 4.7% for all U5M and for those outside the neonatal group (aged 1-59 months) respectively. (4) The proportion of U5M attributable to neonatal conditions has thus increased gradually across the world as seen in Figure 1.2.(2-4)

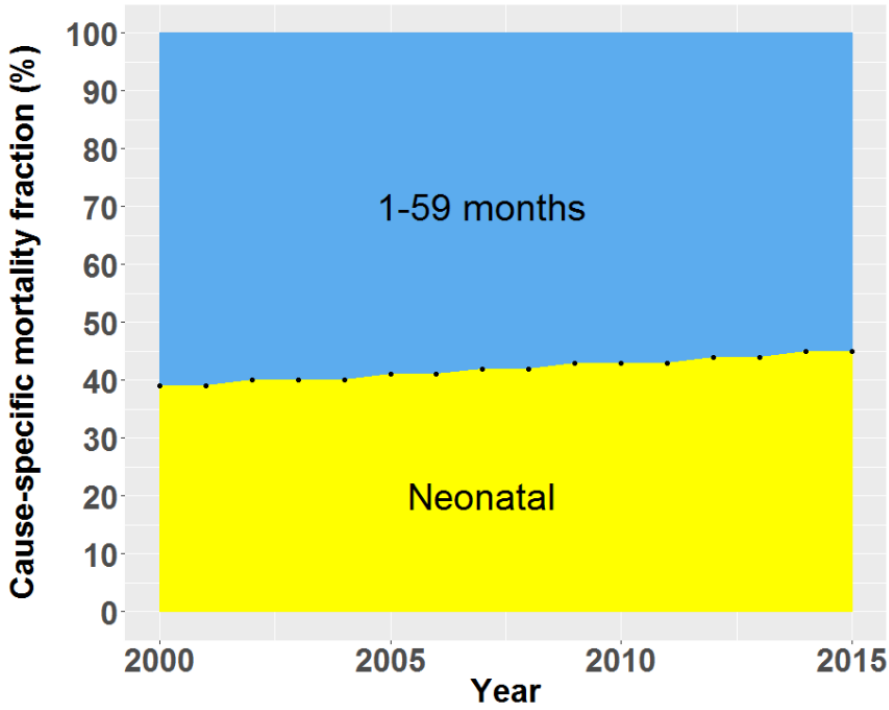



***Figure 1.2:*** *Global under-5 case mortality fractions in 2000-15. Adapted from Liu L Et al. Global, regional, and national causes of under-5 mortality in 2000-15: an updated systematic analysis with implications for the Sustainable Development Goals. The Lancet 2016;388(10063):3027-35*

In 2015, this proportion increased to 45% with 2.7 million neonatal deaths out of a total of 5.9 million deaths in the first five years of life.(4) An estimated 36% of these neonatal deaths occurred on the first day of life and 73% within the first six days.(1) This growing relative burden is also seen in data on cause specific mortality where two of the top three causes of under 5 deaths in 2015 are neonatal: pre-term birth complications (1million deaths) and intrapartum related deaths (0.7 million deaths) respectively.(4)

In the post MDG era (after 2015), the Sustainable Development Goals (SDG) seek to build upon the progress already achieved. Specific to health, SDG 3 aims to "Ensure healthy lives and promote well-being at all ages".(5) However, the insufficient progress in neonatal mortality during the MDG era means that neonates

are a high priority group deserving of specific attention. This is explicitly captured in SDG 3.2 which states “By 2030, end preventable deaths of *newborns* and children under 5 years of age, with all countries aiming to reduce neonatal mortality to at least as low as 12 per 1,000 live births…”.(5) This transition to SDG will require a greater focus on health systems in order to achieve this target for neonatal mortality.(5, 6) Before considering the role of health systems, I first present regional and subnational disparities in neonatal survival in Sections 1.1.3 and 1.1.4.

*1.1.3 Regional variation in neonatal mortality*

The global averages seen in Section 1.1.2 mask considerable disparity in neonatal survival between regions and countries as seen in Tables 1.1 and 1.2 .The highest NMR (29.81 per 1000 live births, South Asia) is more than 10-fold the lowest observed in high income regions (NMR 2.69 per 1000 live births).The risks in South Asia and sub-Saharan African are similar (29.81 and 27.81 per 1000 live births respectively) but are about twice that observed in the next highest region (14.01 per 1000 live births,North Africa and Middle East).

***Table 1.1**: Global variation in estimated neonatal mortality rate for 2015 including the country specific estimate for Kenya.*

| Region | NMR* | Uncertainty interval |
|---|---|---|
| 1. **Global** | 18.62 | 17.62 to 20.14 |
| 2. **High income** | 2.69 | 2.50 to 2.91 |
| 3. **Southern Latin America** | 6.06 | 5.69 to 6.47 |
| 4. **Central Europe,eastern Europe and central Asia** | 7.19 | 6.27 to 8.16 |
| 5. **Southeast Asia,east Asia and Oceania** | 8.16 | 6.84 to 10.12 |
| 6. **Latin America† and Caribbean** | 9.28 | 7.29 to 11.60 |
| 7. **North Africa and Middle East** | 14.01 | 12.33 to 15.77 |
| 8. **Sub-Saharan Africa** | 27.81 | 26.11 to 29.83 |
| 9. **South Asia** | 29.81 | 27.23 to 32.67 |
| **Kenya** | 21.10 | 19.16 to 23.16 |

*NMR: Neonatal mortality rate per 1000 live births
†Central,Andean and Tropical Latin America
Note: Adapted from Wang H, Bhutta ZA, Coates MM et al. Global, regional, national, and selected subnational levels of stillbirths, neonatal, infant, and under-5 mortality, 1980-2013;2015: a systematic analysis for the Global Burden of Disease Study 2015. The Lancet. 2016;388(10053):1725-74

Sub-Saharan Africa saw the highest NMR in the world in 2015; Mali, 40.6 per 1000 live births (Table 1.2). In Kenya, NMR at 21 per 1000 live births is almost half the rate in Mali and two thirds the rate of tenth placed Cameroon.(2)

***Table 1.2**: Ten countries with the highest neonatal mortality rate in sub-Saharan Africa in 2015.*

| Country | NMR* | Uncertainty interval |
|---|---|---|
| 1. **Mali** | 40.60 | 34.58 to 47.11 |
| 2. **Central African Republic** | 40.16 | 32.17 to 49.92 |
| 3. **Lesotho** | 37.47 | 32.64 to 42.98 |
| 4. **Guinea-Bissau** | 36.78 | 31.11 to 42.87 |
| 5. **Côte d'Ivoire** | 36.44 | 32.49 to 40.61 |
| 6. **Chad** | 36.19 | 31.37 to 41.64 |
| 7. **Sierra Leone** | 35.78 | 31.99 to 39.48 |
| 8. **Guinea** | 35.21 | 30.58 to 40.06 |
| 9. **Nigeria** | 34.59 | 30.11 to 39.16 |
| 10. **Cameroon** | 31.64 | 27.18 to 35.92 |
| **Kenya†** | 21.10 | 19.16 to 23.16 |

*NMR: Neonatal mortality rate per 1000 live births
†Kenya included for comparison with the top ten in Africa
Note: Adapted from Wang H, Bhutta ZA, Coates MM et al. Global, regional, national, and selected subnational levels of stillbirths, neonatal, infant, and under-5 mortality, 1980-2013;2015: a systematic analysis for the Global Burden of Disease Study 2015. The Lancet. 2016;388(10053):1725-74

Due to persistently high fertility rates it is expected that Africa will account for most of the observed increase in population worldwide between 2015 and 2100 as shown in Figure 1.3. As a result, the region is expected to maintain the highest numbers of live births per 1000 people in the population over the same period.(7) Consequently, without improvement in neonatal survival rate (per 1000 live births), the absolute number of neonatal deaths is expected to increase.(3)

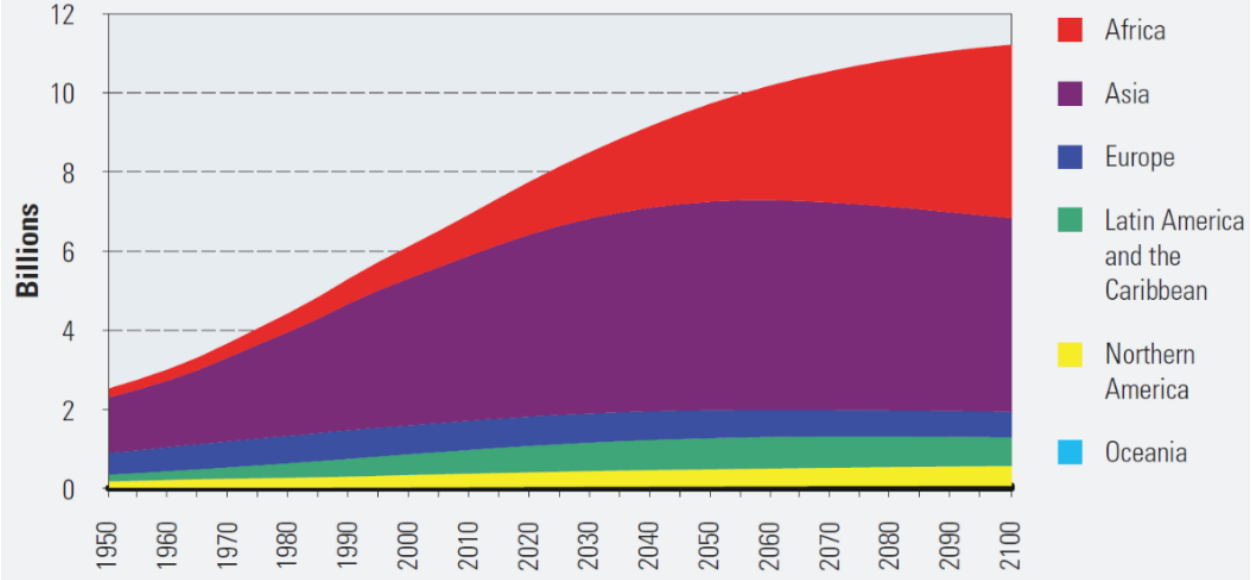



***Figure 1.3:*** *Global population growth by major area.* *Reprinted from United Nations. World Population Prospects: The 2015 Revision, World Population 2015 Wallchart.ST/ESA/SER.A/378*

Despite Kenya being the strongest economy among the five East African Community (EAC) countries there has been relatively slow reduction in NMR between 1990 and 2015; Figure 1.4.(8) Consequently, Kenya is now placed fourth in the EAC.(9)

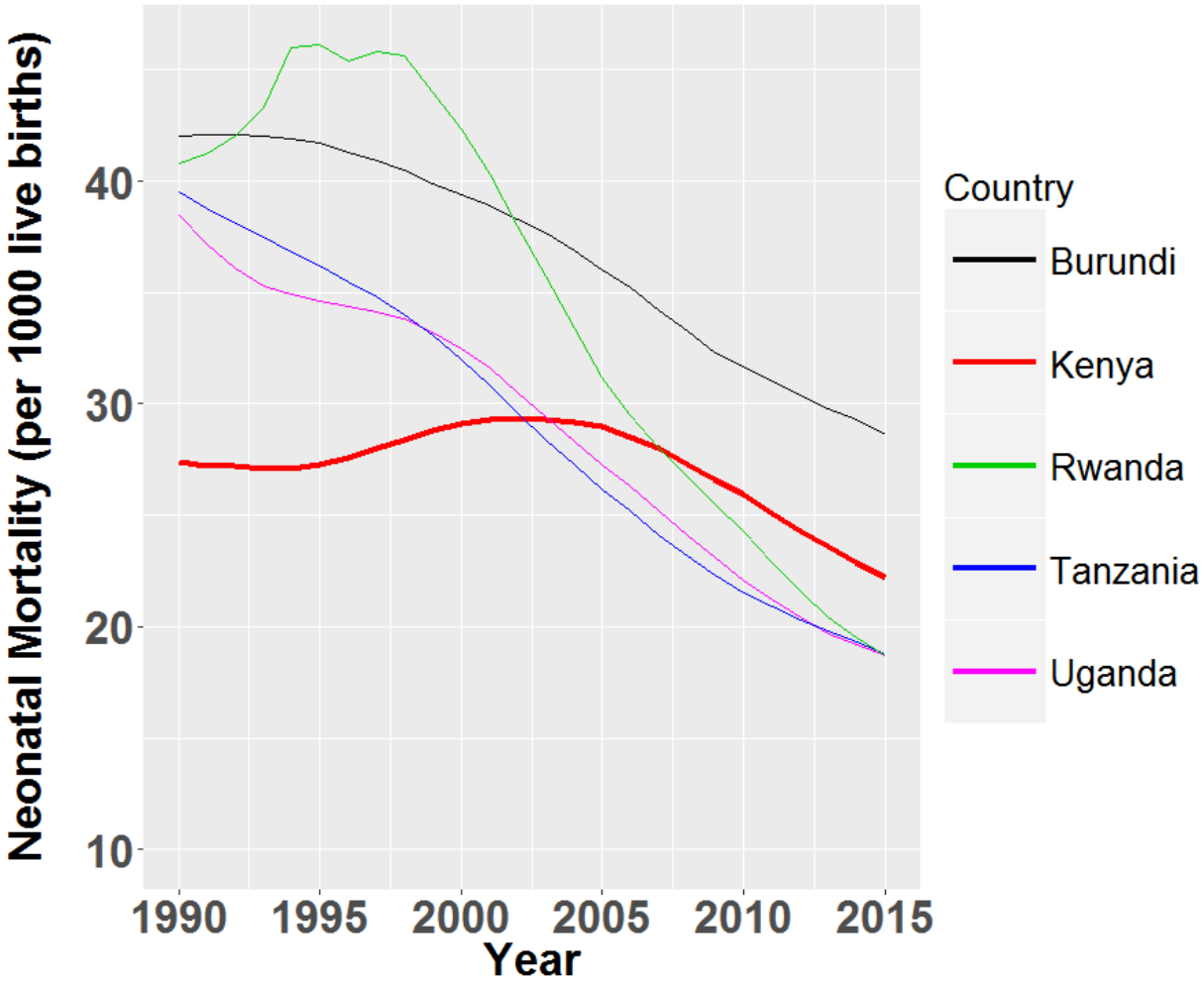



***Figure 1.4:*** *Neonatal mortality rate in East Africa 1990-2015.Adapted from United Nations Children's Fund. Number of neonatal deaths 2015 [updated January 2016].*

The above examples of the disparities in neonatal survival - globally, in Africa and East Africa- demonstrate that despite needs of children being universal, probability of these being met is not.(7) This pattern is replicated at sub-national level as is seen in the following section on the distribution of neonatal deaths in Kenya.

*1.1.4 Population level estimates of neonatal mortality in Kenya*

In Kenya, the proportion of U5M due to neonatal deaths has increased in keeping with the global trend. The proportion in 2015 stands at 45% up from 27% in 2000 as shown in Table 3. With the reduction in total neonatal deaths down to 33,741 in 2015 (Table 3), the country is now placed 15th globally in the absolute numbers of neonatal deaths.(9)

***Table 1.3:*** *Progress in under-5 and neonatal mortality in Kenya 2000-15*

| Year | Under 5 deaths* | Neonatal deaths* | NMR† | NMR/U5M‡ (%) |
|---|---|---|---|---|
| **2000** | 125108 | 33956 | 29.1 | 27 |
| **2005** | 111088 | 38793 | 29.0 | 35 |
| **2010** | 88067 | 37309 | 25.9 | 42 |
| **2015** | 74429 | 33741 | 21.0 | 45 |

* Total (absolute) number of deaths
† Neonatal Mortality Rate (NMR) per 1000 live births
‡ NMR as a proportion of under 5 mortality per thousand live births
Note: Adapted from Liu L, Oza S, Hogan D et al. Global, regional, and national causes of under-5 mortality in 2000-15: an updated systematic analysis with implications for the Sustainable Development Goals. The Lancet. 2016;388(10063):3027-35 and UN Inter-agency Group for Child Mortality Estimation. Child Mortality Estimates: CME Info 2017.

There is also inequity in the risk of neonatal death within the country as shown in Table 1.4. These data are derived from the Kenya Demographic and Health Survey (KDHS) 2014 report published in 2015.(10) The child mortality estimates were computed from women's birth history data for the five years preceding the survey (i.e. birth outcomes for 2009-2014).To arrive at the estimates, the total number of neonatal deaths reported over the five years are expressed as a proportion of total live births in the same period. Out of the eight administrative regions, Nairobi which is the capital city, had by far the highest neonatal mortality in 2009-2014.

***Table 1.4**: Neonatal mortality rate by region in Kenya for 2009-2014.*

| Region | NMR* |
|---|---|
| **Nairobi** | 39 |
| **Coast** | 25 |
| **North Eastern** | 24 |
| **Eastern** | 24 |
| **Central** | 24 |
| **Rift Valley** | 20 |
| **Western** | 19 |
| **Nyanza** | 19 |

*Neonatal Mortality Rate (NMR) in the 10-year period preceding the 2014 Demographic and Health Survey
Note: Adapted from Kenya Demographic and Health Survey 2014.

This within country inequity in health status has been recognised as a key contributor to the persistently high child mortality in developing countries and therefore deserves appropriate attention.(11) In Kenya, Nairobi is therefore a priority region in sub-national efforts to improve neonatal survival and by extension national equity in health. Health systems and especially hospitals are expected to play a big role in these efforts. In the next section (1.2), I will therefore look at the role of hospitals in improving neonatal survival.

## 1.2 Interventions to improve neonatal survival and the role of hospital care

### *1.2.1 Essential neonatal interventions*

Essential newborn care has been defined as “a comprehensive strategy designed to improve the health of newborns through interventions before conception, during pregnancy, and soon after birth, and in the postnatal period”.(12) These critical time periods for intervention are illustrated in Figure 1.1. This approach recognises that maternal and neonatal health are inextricably linked and interventions are best delivered in an integrated continuum of care.(13)

The recommended essential interventions for newborns are based on the best currently available evidence and these are included in the World Health Organisation (WHO) guidelines.(14-16) In Kenya, these guidelines have been adapted to suit the local context and adopted by the Ministry of Health as the standard of care.(17, 18). In addition to clinical effectiveness, there is evidence demonstrating cost effectiveness.(15, 19) Effectiveness of these interventions will however only be realised in the presence of appropriate service delivery platforms in the health system.

### *1.2.2 Hospitals as a service delivery platform for essential neonatal interventions*

As neonatal mortality dips below 30 per 1000 births, basic hospital care (essential newborn care) will be required to improve survival further.(13, 14) Essential interventions include antibiotics for presumed severe bacterial infection, oxygen for respiratory support, safe provision of intravenous fluids and feeding support. These target the common problems of small and sick newborns e.g. complications of preterm birth, infections and intrapartum events (occurring during the process of

birth). To this end, scaling up of hospital services for neonates is required in low and middle income countries (LMICs) where most neonatal deaths occur.(20-22) Increasing the coverage and quality of maternal and essential neonatal care in these countries with high mortality rates could prevent up to 71% of neonatal deaths with 82% of this effect attributable to facility based care.(15) Facility based care has indeed been shown to reduce neonatal mortality.(23)

In Kenya, facility based delivery is increasing partly as a result of the free maternity care programme rolled out by the national government on 1st June 2013.(10, 24) It is therefore expected that there will be increased demand for care for small and sick neonates in hospitals. The continued increase in demand for neonatal care as a proportion of all paediatric care in hospitals is expected to accelerate an existing trend. In one hospital in rural Kenya, neonatal admissions as a proportion of under 5 admissions increased from 11% to 20% as overall childhood illness declined over a 19-year period.(25) This increase in demand must be matched with quality service delivery for hospitals to have an impact on survival.

*1.2.3 Bottle-necks to the delivery of quality essential neonatal care in hospitals.*

Hospital care for neonates in low and middle income countries is however beset with quality challenges.(26-29) Essential interventions facing the greatest health system bottlenecks to delivery at scale are: (1) prevention and management of preterm births, (2) inpatient supportive care of ill and small newborns, (3) management of severe infections and kangaroo mother care.(27, 28) An assessment of these bottlenecks by stakeholders in eight countries with a high burden of neonatal mortality (including Kenya) identified the health workforce and

service delivery as the most important constraints to delivery of essential maternal and neonatal care.(27) The key issues pertaining to the health workforce were shortage of workers, lack of skills and low competency whilst in service delivery, poor quality of care was the major issue.(27)

In Kenya, poor quality of essential neonatal service delivery in hospitals has indeed been observed.(30-35) This includes errors in dosages for antibiotic prescriptions.(32) Besides the potential for harm from over dosage, with under dosage there is concern for insufficient treatment of infections with most deaths from severe neonatal infections occurring in developing countries like Kenya.(36, 37) Other indicators of poor care observed include: unavailability of resources such as pulse oximeters (for measuring oxygen content of blood), poor documentation of symptoms of severe illness and inappropriately prescribed volumes of intravenous fluids and feeds.(32) Addressing quality of care must therefore be a part of efforts to improve coverage and utilisation of essential neonatal care as illustrated in Figure 1.5.(38)

**The image originally presented here cannot be made freely available via ORA because of copyright. The image was sourced from** *Crockett M, Avery L, Blanchard J. Program science—a framework for improving global maternal, newborn, and child health. JAMA Pediatrics. 2015;169(4):305-6*

***Figure 1.5:*** *Conceptual framework for improving Maternal, Newborn and Child Health (MNCH) across the continuum of care. Reprinted from Crockett M, Avery L, Blanchard J. Program science—a framework for improving global maternal, newborn, and child health. JAMA Pediatrics. 2015;169(4):305-6*

To improve survival and deliver quality care in LMICs, rationally designed systems and services will be therefore required to achieve the goal of universal health coverage with essential neonatal care.(26, 39) These systems must address the key bottlenecks in the health workforce and service delivery. Proposed solutions include; for the health workforce - development of policies for staff deployment and recruitment, for quality of care - development of standards of care and clinical practice guidelines, audits and regular review of performance in all health facilities to inform quality improvement.(27)

One strategy employed in implementing such solutions for maternal and neonatal health in high income regions is the regionalisation of perinatal care. This approach aims to improve patient outcomes by optimising capabilities to care for different illnesses in different facilities. Scarce resources for example specialised neonatal physicians and nurses are concentrated in one or a few specialised units while basic and essential care is offered in a larger number of lower level units to ensure

access. Balancing is thus achieved between centralisation and access. Regionalisation therefore takes into account that economic constraints preclude the ability of all hospitals to maintain the sufficient staff and equipment to care for all medical conditions.(40) These economic constraints are more pronounced in lower income countries and such approaches may therefore contribute towards rational distribution of scarce skilled staff and medical supplies and co-ordination of care within regions based on patient needs. Concentrating resources for the management of complex conditions at specialised hospitals may indeed improve outcomes.(40)

An example in Kenya would be regionalisation of neonatal care in Nairobi which has the highest neonatal mortality risk and is the largest county by population in Kenya. This is consistent with the Kenyan Ministry of Health's National Referral Strategy where "organisation of service delivery into six levels of care is intended to rationalise the delivery of health services within the health system for efficient use of existing resources".(41) Successful linkage between the levels requires a functional referral system the absence of which is currently a constraint to service delivery in Kenya.(27) Such rational design of services in Nairobi will entail the use of information to support decision making processes.(42) This in turn shall require good quality information and decision support tools as presented in the next section.(42, 43)

## 1.3 Use of information in decision making to improve quality of hospital care

### *1.3.1 The role of health data in decision making processes*

The inclusion of information as one of the WHO six health system building blocks emphasises its indispensable role in health systems.(44) Ensuring the production, analysis, dissemination and use of reliable and timely health information is necessary if improved health is to be achieved.(44) For this reason, to improve neonatal survival in Kenya through better hospital care, better clinical information is required.(45-48) Where information is available, specific efforts will be required to facilitate its use as Nutley and colleagues state “The lack of demand for and use of data limits the health system’s ability to respond to priority needs throughout its many levels”.(49) They identify activities that will strengthen demand for and use of data such as improvement of data availability and quality in order to realise the use of data and information in decision making processes. Beyond availability and quality, the incorporation of data in decision making processes is also influenced by the mode of presentation.(50)

### *1.3.2 Decision making processes in hospital care for neonates in Kenya*

Decision making processes to improve hospital care for neonates in Kenya may be seen to occur at three levels as shown in Table 1.5. These are: (1) individual patients, (2) groups of patients and (3) the planning and organisation of service delivery.

***Table 1.5**: Illustration of how information can guide decision making processes in hospital care for neonates in Kenya*

| Level | Example questions | Current decision support approach in Kenya | Application | Limitations |
|---|---|---|---|---|
| 1.Individual patient | How sick is the patient? | Individual signs and symptoms of severe illness (51) | Treatment decisions | Prioritises sensitivity at the expense of specificity(51) |
| 2.Groups of patients | How similar are groups of patients and their outcomes? | Gestation and birthweight categories (52) | Fair comparison of mortality across place and time | Residual variability within categories after adjustment |
| 3.Service delivery | How many nurses are required? | Experience ,expert opinion(53) | Planning service delivery | Optimality *(54)may not be achieved |

* Optimality: the most advantageous balancing of costs and benefit

*a) Decision making at individual patient level*

At the individual patient level, a measure of the severity of illness is useful to guide treatment plans. In Kenya, the national paediatric clinical practice guidelines recommends a set of nine clinical signs that are predictors of need for hospital level care for severe neonatal disease e.g. severe bacterial infection. The signs are: change in level of activity, convulsions, feeding difficulty, abnormal temperature, fast breathing, severe chest wall indrawing, grunting, cyanosis (or low oxygen saturation) and bulging fontanelle.(17, 51) These are intended for use in clinical settings where health workers have limited specific neonatal expertise and limited diagnostic support e.g. pulse oximetry to determine blood oxygen levels and blood cultures to identify bacterial causes of severe infection. These conditions apply to Kenyan first referral level hospitals where junior staff offer most care and diagnostic facilities are largely absent.(51) The presence of any one of these nine clinical signs indicates the need for hospital admission and initiation of empirical antibiotic therapy for individual patients. However, this screening approach prioritises sensitivity over specificity and leads to antibiotic treatment in those without

bacterial infection.(51) A more detailed description of these clinical signs is provided in Chapter 6.

*b) Decision making on groups of patients*

Classification of groups of neonates into categories based on risk of an outcome (e.g. death) is useful in comparing performance of facilities and in clinical research. Such comparisons of clinical outcomes amongst facilities providing in-patient neonatal care requires taking into account differences in the risk profiles of patients (case-mix adjustment).(55) In research, such categorisation is useful for instance in balancing the allocation to treatment and control groups in clinical trials (risk-stratification). In Kenyan hospitals, birth weight and gestational age are typically used to define risk categories for neonates: Tables 1.6 and 1.7.

***Table 1.6**: Birth weight categories*

| Category | Weight (g) |
|---|---|
| **Macrosomia** | >4000 |
| **Normal** | 2500 to 4000 |
| **Low birth weight** | 1500 to 2499 |
| **Very low birth weight** | 1000 to 1499 |
| **Extremely low birth weight** | <1000 |

***Table 1.7:** Gestational age categories*

| Category | Gestation (weeks) |
|---|---|
| **Postdates** | >42 |
| **Term (Normal)** | 37 to 42 |
| **Late pre-term** | 34-36 |
| **Moderate preterm** | 32-33 |
| **Very pre-term** | 28-31 |
| **Extremely pre-term** | <28 |

Weight and gestational age outside the normal category is associated with increased risk of morbidity and mortality. There is in particular a strong association between low birthweight (<2500g) and pre-term birth (<37 weeks) and risk of

mortality.(56) In Kenyan hospitals, birthweight has the advantage of availability. A survey of 22 neonatal units revealed good documentation of birth weight (93%) compared with gestational age (55%).(57) However, there is residual variability in outcome within birth weight categories. Low birth weight for instance, results from restriction of foetal growth (intra-uterine growth restriction-IUGR), preterm birth or a combination of the two.(56) Babies who are both preterm and experienced IUGR have a higher risk of death compared with preterm or IUGR alone.(58) Even where gestational age is available, there is heterogeneity within categories as recent work has demonstrated 12 clinical sub-groups (phenotypes) related to preterm birth.(59)

The use of birthweight and gestational age from routine hospital data in Kenya to describe groups of neonates is currently limited. Such data should be available for all patients but even if this were the case, it still leaves the problem of residual variability in the established categories (Table 1.6 and 1.7). In light of this, the use of other clinical characteristics documented in the routine patient records may be considered to augment birth weight in efforts to achieve effective case-mix adjustment.

*c) Decision making for service delivery planning and organisation*

Service delivery planning and organisation is essential to ensure available resources are transformed into improved health outcomes. This as noted by The World Bank, requires organisations (health facilities) to “combine financial, physical and human resources to deliver health services”.(60) Nurses are the back bone of in-patient neonatal care and low nurse to patient ratio may be associated with increased mortality.(61-63) This is because effective delivery of life-saving

interventions such as intravenous antibiotics, safe intravenous fluids and enteral feeding requires labour intensive input from skilled nursing staff. Rational allocation of nurses for neonatal care is therefore critical in Kenya(Table 1.5); a country faced with a shortage of nurses.(64)

A key component of rational investment in human resources for health - such as nurses - is the alignment of investment with the population health needs. Reduction of neonatal mortality will require a scaling up of in-patient hospital care which is dependent on nursing care. Availability of nurses skilled in the provision of neonatal care is therefore a clear population health need in Kenya. The use of patient needs to determine nursing workload is further discussed in Chapter 3. Resource allocation in the absence of such patient data may lead to failure to achieve the most advantageous balancing of costs and benefit (optimality).(54)

Better data information and data can thus inform and thereby improve decision making for individual patients, groups of patients and in the planning and organisation of service delivery. I shall now discuss the potential role of a specific type of data-prognostic data- in these three decision making processes.

### 1.4. Role of prognostic data in decision making to improve quality of health care

The purpose of health care is to improve health related outcomes (prognosis) for example in reducing mortality. However therapeutic decision making is traditionally based on the likelihood of a certain condition (diagnosis) rather than on prognosis.(65) In the absence of diagnostic tests (e.g. blood cultures for severe infections) as is common in LMICs, clinical decision making relies upon prognostic data such as clinical signs which are predictive of mortality. Nonetheless, prognostic data can augment diagnostic decision making by for instance providing information on who is likely to respond to certain treatments.(66)

The **PROG**nosis **RES**earch **S**trategy (PROGRESS) partnership is a U.K Medical Research Council funded collaboration "developing understanding in research into quality of care outcomes, prognostic factors, risk prediction models, and predictors of differential treatment response" .(67) PROGRESS organises prognostic research into four themes:

i. Fundamental prognosis; describing and explaining future outcomes in people with a disease or health condition in relation to current diagnostic and treatment practices.(68)
ii. Prognostic factor research; identifying factors associated with a subsequent clinical outcome in people with a particular disease or health condition.(69)
iii. Clinical prediction modelling; combining multiple predictors to calculate the risk of a particular outcome for individual patients.(70)

iv. Stratified medicine; investigating the targeting of treatments or interventions according to the biological or risk characteristics shared by subgroups of patients.(71)

My work applies this framework with a focus on neonates in Kenya as this approach has not been used hitherto. As illustrated in Figure 1.5, better prognostic data for neonates receiving hospital care could be used to support decision making processes leading to better quality of care and improved survival.

## 1.5 Prognosis after neonatal admission in Kenyan hospitals

### *1.5.1 Review of prognosis data from Kenyan hospitals*

There is no national database on survival of neonates admitted to neonatal units in Kenyan hospitals. This is a reflection of data quality challenges in the current hospital information systems and the national vital registration system.(46, 72) As a result, there are few data available from these neonatal units.I conducted an electronic search in PubMed and African Journals Online (AJOL) for any studies reporting mortality during hospital care for neonates in Kenya over a ten year period (2007-2016).Key words used were “neonate” ,“ newborn” and “Kenya” including the medical subject headlines (MeSH) terms where applicable (Appendix 8). This yielded 15 relevant studies (Table 1.8).

***Table 1.8:*** *Neonatal inpatient case fatality in Kenyan hospitals*

| Study | Year | Hospitals | Sample | Weight* | CFR† |
|---|---|---|---|---|---|
| **Were et al[73]** | 2007 | 1 | 344 | <1500g | 159/344 (46%) |
| **Mwaniki et al[74]‡** | 2009 | 1 | 1105 | All | 336/1106 (30%) |
| **Were et al[75]** | 2009 | 1 | 260 | <1500g | 116/260 (45%) |
| **Talbert et al[24]‡** | 2010 | 1 | Unspecified | All<br>All | 24%§ vs 21%‖ (0-6 days)<br>8%§ vs 4%‖ (7-28 days) |
| **Mwaniki et al[76]‡** | 2010 | 1 | 1572¶ | All | 300/1572 (19%) |
| **Mwaniki et al[25]‡** | 2010 | 1 | 8756 | All | 2053/8756 (24%)** |
| **Mwaniki et al [77]‡** | 2010 | 1 | 5114 | All | 1011/5114 (20%) |
| **Marete et al[78]** | 2011 | 1 | 135 | <2500g | 62/135 (46%) |
| **Gathara et al [31]** | 2011 | 8 | 798 | All | 241/639 (38%) |
| **Kohli-Kochhar[79]** | 2011 | 1 | 152 | Unspecified | 4/152 (3%)†† |
| **Yego et al[80]** | 2013 | 1 | 200 | Unspecified | 68/1000 live births |
| **Aluvaala et al [32]** | 2015 | 22 | 1065 | All | 180/1065 (17%) |
| **Ibinda et al[81]‡** | 2015 | 1 | 191 | All | 118/191 (62%) ‡‡ |
| **Aluvaala et al [57]** | 2015 | 5 | 1384 | All | 263/1384 (19%) |
| **Myhre et al [250]** | 2016 | 1 | 118 | <2500g | 10/46(22%) vs 7/72 (11%)# |

* Birth weight eligible for inclusion in the study
† CFR: Case Fatality Rate
‡ Same facility, different studies, part of a demographic surveillance system
§ All outborn, numbers not provided
‖ All inborn, numbers not provided
¶ Total neonatal admissions were 3302: excluded 1702 from outside DSS area and 28 readmissions
** Average over 19 years (1990-2008), reduced from 30.8% in 1990 to 16.5% in 2008
†† Patients with blood culture positive sepsis only in a private tertiary hospital
‡‡ Neonatal Tetanus only
# Before (22%) and after (11%) introduction of Continuous Positive Airway Pressure (CPAP) respectively

Six of these studies are from one rural facility in coastal Kenya affiliated to the KEMRI-Wellcome Trust research programme.(24, 25, 74, 76, 77, 81) Three are from two national referral hospitals.(73, 75, 80) Heterogeneous populations were included with differences in age at admission and birth weight which precludes direct comparison. None of the studies conducted time to event analysis and therefore there is no data on probability of death at different time points during the admission episode. More data is however expected from facilities providing in-patient neonatal care in Nairobi.(82)

The few published studies illustrate the scarcity of health facility prognostic data for neonates in Kenya relative to the population (>40 million), number of health

care facilities (≈8,000), annual birth cohort in 2016 (≈1.6 million) and high neonatal mortality. Where such data is available, it has contributed to better care as illustrated in the examples that follow.

*1.5.2 Examples of use of neonatal prognostic data in decision making*

Where available (predominantly from high income countries), prognostic data have informed clinical and policy level decision making in addition to guiding research on interventions.(68) At policy level, the recognition that very low birth weight babies have better survival in high volume neonatal units has influenced the regionalisation of perinatal care in high income regions.(83, 84) The Kenya national paediatric clinical practice guideline recommends that babies weighing less than 1500g should not receive enteral feeding on the first day of life as this increases the subsequent risk of developing Necrotising Enterocolitis; an intestinal condition associated with high morbidity and mortality.(17, 85) This is an example of how despite the absence of local data, the clinical understanding of prognosis has influenced clinical policy. Finally in research, therapeutic trials of body cooling after perinatal asphyxia included infants aged less than six hours as evidence suggests that effectiveness of cooling declines after this time.(86-88) In light of these examples, what then is the impact of the paucity of prognostic data on neonatal care in LMICs?

*Effect of lack of neonatal prognostic data from LMICs*

Poor availability of neonatal mortality data (fundamental prognosis) in LMICs necessitates reliance on estimates based on statistical modelling. (2, 89, 90) Whereas the resultant estimates are useful at a global and national level for advocacy and priority setting, they have limited utility for decision making at sub-national levels e.g. organisation of perinatal care in a region like Nairobi in Kenya and informing guidelines for the deployment of new technologies such as bubble continuous positive airway pressure (BCPAP). Consequently, clinical and policy decisions at sub-national levels are made amidst considerable uncertainty. This may contribute to failures of care delivery and coordination, overtreatment, administrative complexity and pricing failures.(91) All of these would prevent hospitals from achieving their potential in the reduction of neonatal mortality in Kenya.

In this Section (1.5) I have focused on the scarcity of prognostic data and its impact in LMICs under the first PROGRESS framework theme i.e. Fundamental prognosis as described Section 1.4 (page 19).(68) I will now explore the role of the third PROGRESS theme-clinical prediction modelling-in neonatal medicine.

## 1.6 Role of clinical prediction models in supporting decision making

### *1.6.1 Overview of clinical prediction models*

Clinical prediction models support decision making by combining multiple patient characteristics to calculate the probability that a specific disease or condition is present (diagnostic models) or that a specific event will occur in the future (prognostic models) as seen in Figure 1.6.(70, 92) Collectively diagnostic and prognostic models are referred to as clinical prediction models.

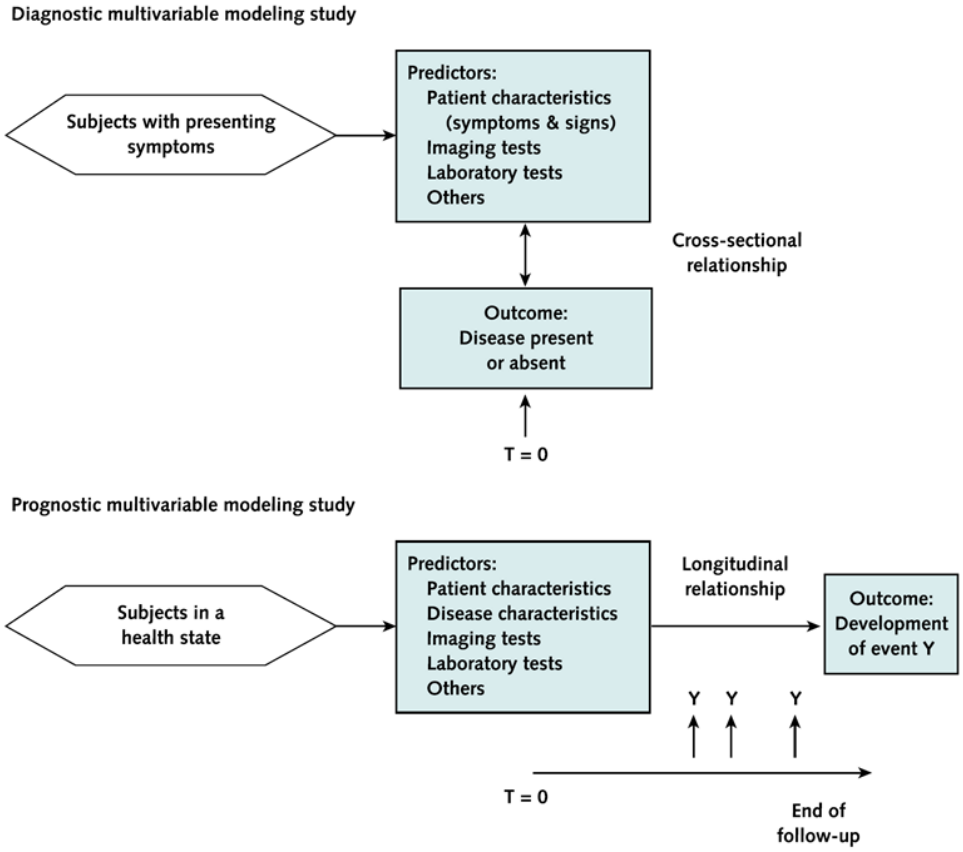



***Figure 1.6**: Types of clinical prediction models. Reproduced from Collins GS, Reitsma JB, Altman DG et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): The TRIPOD Statement. Annals of Internal Medicine. 2015;162(1):55-63.*

In addition, they may estimate resource use to inform service delivery planning.(93, 94) They have variably been termed as prediction rules, probability assessments, prediction models, decision rules and risk scores.(95)

*1.6.2 Development of prognostic models*

The recognition of the potential utility of prediction models has seen a rapid increase in publications in the recent past as seen in Figure 1.7. I conducted this illustrative search on PubMed using the key words “prediction”, “model”,” score”, “tool”,” index “,” rule” for the period 1970-2016. There were 10 publications in 1970 increasing to 8838 in 2016.

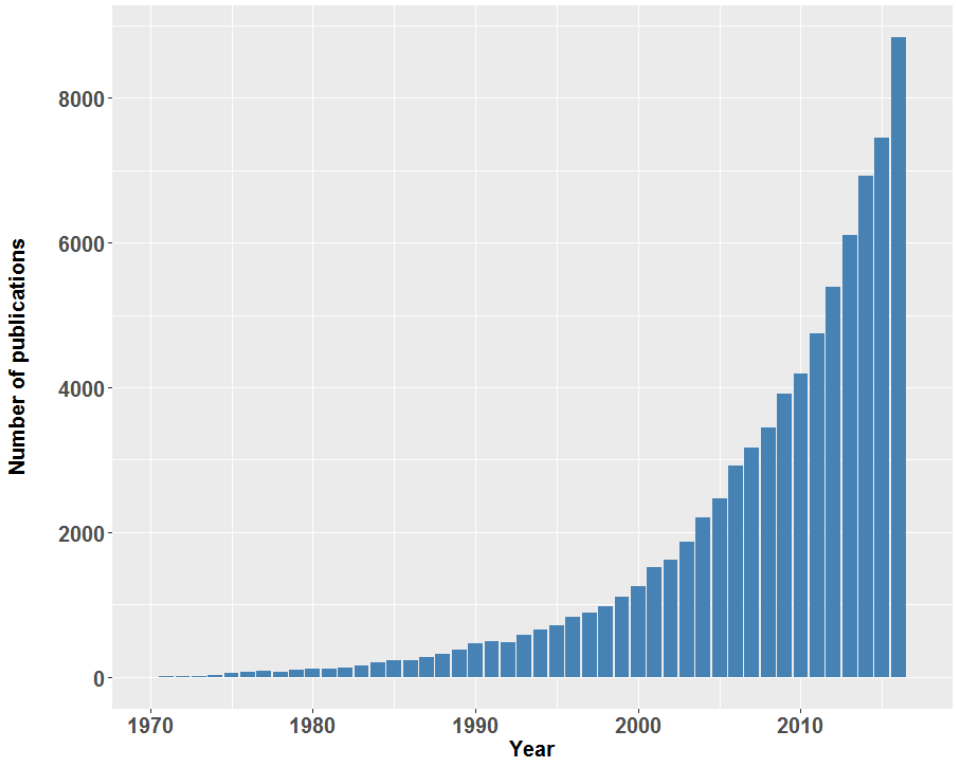



***Figure 1.7:*** *Illustrative search of prediction models indexed on PubMed 1970-2016.*

Numerous reviews of prediction modelling studies have however revealed many models have serious limitations in their design, conduct and reporting.(96-99) In light of these limitations, guidance to inform model development, reporting of primary studies and conduct of systematic reviews have been published.(92, 100, 101).

The development of prediction models may be described in four phases: development, validation, impact assessment and implementation as illustrated in Figure 1.8.(95)

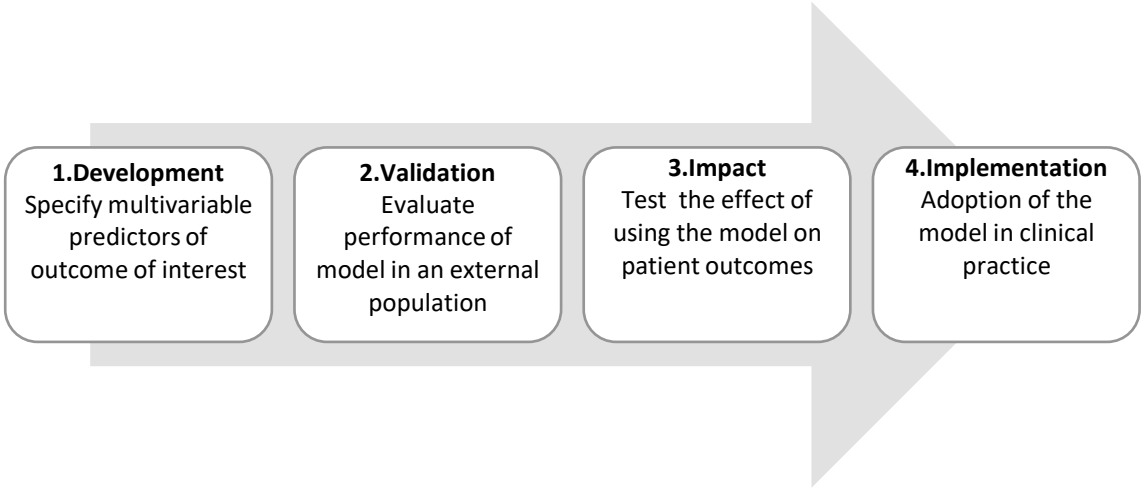



***Figure 1.8:*** *Phases in the development of prediction models. Adapted from Adams ST, Leveson SH. Clinical prediction rules. BMJ. 2012;344: d8312*

In the development stage, multiple variables or predictors are identified which are then entered into a statistical model to predict the outcome of interest.(70) These variables are patient characteristics measured at a defined time point when the model is intended to be used e.g. birthweight, gestational age, arterial oxygen saturation. A statistical model is then specified such as a logistic regression model, to compute the probability of an outcome, e.g. death in a patient with a given combination of these variables. An example model expressed as an equation is given in Box 1.1.

| Predicted risk of death = Birthweight +gestational age + arterial oxygen saturation |
|---|

***Box 1.1:*** *An example of a prediction model formula*

Validation is aimed at evaluating the performance of the model in a different population. This phase evaluates the discrimination ability and calibration of the model. Model discrimination refers to the ability of the model to differentiate between patients with different outcomes e.g. dead vs alive whilst calibration is the agreement between the predicted and observed risks.(70) The impact of using a validated model in the general population is best assessed in a (cluster) randomised controlled trial by comparing outcomes (e.g. patient outcomes, decision-making, costs of care) in two groups; one where usual care is provided and in the other where model predictions are provided to guide clinical decisions.(70) Finally, the implementation phase promotes adoption of the model into routine clinical practice by for example, incorporating it in existing electronic health record systems.(70) These phases are described in detail in the methods section presented in Chapter 2.

*1.6.3 Overview of prediction models in neonatal medicine*

Prediction models have been developed and used widely in neonatal intensive care in high income countries.(102) Applications include using predicted risks to provide prognostic information to parents and identification of high risk infants eligible for particular clinical trials.(102) In-depth discussion of neonatal prediction models is presented in Chapter 3.

These models predominantly use physiological parameters measured in neonatal intensive care and include gestational age to predict survival data which are frequently missing in medical records of low-income countries(LIC).(32, 57, 103) In contrast, the neonatal treatment intensity scoring system (NTISS) uses treatments to predict prognosis and estimate resource use.(102, 104) Using treatments as a modeling approach is of particular interest in Kenya given the information gaps at patient and service delivery planning levels (Table 1.5) and local experience that treatment data abstracted from routine hospital records can be obtained.(30-32) Better survival data coupled with the treatment intensity model approach could therefore provide information for decision making at patient level in addition to informing resource allocation (Table 1.9 and Chapter 3).

***Table 1.9:*** *Potential application of prediction models in neonatal medicine in Kenya*

| Level | Example questions | Role of prediction model | Application |
|---|---|---|---|
| 1.Individual patient | How sick is the patient? | Predict outcome | • Treatment decisions based on probability of outcome(70) |
| 2.Groups of patients | How similar are groups of patients and their outcomes? | Group patients by probability of outcome | • Fair comparison of mortality across place and time(70)<br>• Stratification in clinical trials(70) |
| 3.Service delivery | How many nurses are required? | Estimate resource use | • Planning service delivery(102) |

Table 1.9 illustrates how clinical prediction models could mitigate the limitations of current decision making processes outlined in Table 1.5. For individual patients the predicted probability of death may be used to determine the need for admission and when in-patient care is required, the appropriate level of facility for example district hospital or tertiary referral hospital. This approach could reduce the number of admissions and help refer patients to facilities best suited for their clinical needs. However, this will require the use of locally relevant, well performing and validated prediction models.

## 1.7 Research objectives

### *1.7.1 General objective*

To describe neonatal survival in a large urban low resource setting Kenyan hospital and to explore the potential of clinical prediction models to predict mortality and support decision making.

### *1.7.2 Specific objectives*

For a population of neonates in a low resource setting hospital, I will:

1. Describe the clinical characteristics and analyse in-hospital survival.
2. Develop an essential neonatal treatment intensity score in two steps by:
    a) Conducting a systematic review of the evidence on the performance of existing neonatal treatment intensity scores and assess their quality and applicability to essential neonatal care
    b) Developing and validating an essential treatment intensity model to predict mortality
3. Develop and validate a symptoms and signs model for prediction of mortality based on clinical signs currently recommended for assessment of severity of illness in the Kenyan paediatric clinical practice guidelines.
4. Compare the performance of the treatment intensity score against that of the signs and symptoms score and explore their clinical utility.

## 1.8 Justification

The neonatal mortality rate in Kenya for 2015 (21 per 1000 live births) is a long way off the Every Newborn action plan target of 12 per 1,000 live births or lower by 2030. To achieve this target, within country inequities must be addressed. Further, implementation at scale of quality essential newborn care at hospital level will be required. Currently, the quality of hospital care is wanting and one way to improve it is to promote rational service design using prognostic information to support decisions such as assigning level of care and referral. This will however first require the availability of better quality data on hospital care for neonates in Kenya. Nairobi County has the highest NMR in Kenya. This research project therefore seeks to generate better prognostic data from the busiest neonatal unit in Nairobi County and thereafter use it to explore the performance of locally relevant prediction models and their potential to support decision making. Such tools could have wider value in Kenya and other LMICs.

## Chapter 2 : Methods

### 2.1 Introduction

In this Chapter, I provide a general description of the methods while their specific application is presented in the relevant results chapters. In Section 1.7 of Chapter 1, the four specific study objectives were outlined. In addition, reference was made to the Prognosis Research Strategy (PROGRESS) themes in Section 1.4.(68) In Chapter 2, I will now describe how the PROGRESS approach was applied as an organising framework towards the achievement of the study objectives. This is a novel application of PROGRESS framework in the understanding of prognosis in neonatal medicine (particularly hospital based essential neonatal care in a low resource setting) and is summarised in Table 2.1.

***Table 2.1 :*** *Research objectives under the PROGRESS prognostic research themes*

| Objective | PROGRESS* Theme |
|---|---|
| **1.Describe clinical characteristics and analyse in-hospital survival** | Fundamental prognosis research(68) |
| **2.Develop an essential neonatal treatment intensity score in two steps:**<br>a) Conduct a systematic review evaluating the quality of existing neonatal treatment intensity scores<br>b) Develop and validate a treatment intensity model to predict in-hospital mortality | Prognostic model research(70) |
| **3.Develop and validate a symptoms and signs model to predict in-hospital mortality** | Prognostic model research(70) |
| **4.Compare the clinical utility of the treatment intensity score against that of the signs and symptoms score** | Prognostic model research(70) |

*PROGRESS: **Prog**nosis **Res**earch **S**trategy(68)

As shown in Table 2.1, this work was informed principally by two of the PROGRESS themes; Fundamental prognosis research and Prognostic model research.(68, 70)

Whereas the systematic review of neonatal treatment intensity scores is part of objective 2 - specifically 2a - I present the review methods first (Section 2.2) because it was the first piece of completed work of the DPhil. This systematic review informed subsequent work in objectives 2b,3 and 4 (Table 2.1). In addition, the same data were used for objectives 1, 2b,3 and 4 therefore I describe the study site, data management and data sets (Sections 2.3 and 2.4) before the objective specific methods (Sections 2.5-2.8). Ethical considerations are presented last in Section 2.9.

## 2.2 Systematic review of prediction modelling studies

### *2.2.1 The need for systematic reviews to inform prediction modelling studies*

As I demonstrated in Chapter 1 (Figure 1.7) there has been a proliferation of prediction models in the recent past. Numerous studies have however identified limitations in the published literature including: unclear study designs, dichotomization of candidate predictors and failure to report follow up times, sample size calculations and performance measures.(93, 96-99, 105) In addition, the evaluation of model performance in populations different from that in which it was developed (external validation) is vital before use in clinical practice. However, only a small proportion of models are externally validated.(106) To avoid waste, applicability of existing models should therefore be assessed before developing new ones.(106, 107) In Chapter 1 (Section 1.6.3), we saw that models using treatments as predictors (treatment intensity scores) are of interest for neonatal care in Kenya as they may be used to predict outcome and estimate resource use. For these reasons, I therefore sought to review existing neonatal treatment scores before considering development of a new one.

The **CH**ecklist for critical **A**ppraisal and data extraction for systematic **R**eviews of prediction **M**odelling **S**tudies (CHARMS) provides guidance to identify and critically appraise existing models.(101) The **G**rading of **R**ecommendations **A**ssessment, **D**evelopment, and **E**valuation (GRADE) system is a transparent approach to the assessment of the quality of evidence and its application in the making of recommendations.(108) A general description of the CHARMS checklist and GRADE is provided here in Sections 2.2.2 to 2.2.4 whilst specific application to objective 2a (Table 2.1) is described in Chapter 3.

### 2.2.2 *The checklist for critical appraisal and data extraction for systematic reviews of prediction modelling Studies (CHARMS)*

The CHARMS checklist was developed by the Cochrane Prognosis Methods Group and has two main parts. The first part deals with items to consider in the formulation of the research question (section 2.2.2(a)) and the second data extraction and critical appraisal (section 2.2.2(b)) A summary of these is provided in Table 2.2 and Table 2.3 respectively while the complete checklist is included in Appendix 1.(101)

*a) Formulation of the research question*

The key issues in the formulation of the research question for a systematic review of prediction modelling studies include: defining the population of interest, type of prediction models to be considered and the outcomes of interest. These and other factors are listed in Table 2.2 alongside their application to my systematic review of neonatal treatment intensity scores. In this table, I illustrate how these factors were applied in the formulation of the review question, inclusion and exclusion criteria and the search strategy.

In summary, the scope of the research question was limited to prognostic models that use a treatment intensity scoring approach to predict clinical outcomes and measure resource utilisation for neonates admitted to neonatal units. A broad search strategy (details in Chapter 3) was subsequently designed to maximize on the number of studies identified because this was an area likely to have few published reports.

***Table 2.2:*** *Key items in the design of systematic reviews of prediction modelling studies.*

| Item | Description | Review of neonatal treatment intensity scores |
|---|---|---|
| **1.Prognostic versus diagnostic** | Is the aim to review models that predict future events (prognostic) or current disease status (diagnostic)? | Prognostic models |
| **2.Intended scope of the review** | Define the scope by specifying the intended purpose of the models to be included | Predict clinical outcomes and measure resource utilisation |
| **3.Type of studies** | Specify type:<br>• Development only<br>• Development and external validation<br>• External validation with or without updating | All types included |
| **4.Target population** | Specify the target population relevant to the scope of the review | Neonates admitted to a neonatal unit in any country |
| **5.Outcome to be predicted** | Specify the outcome of interest to be predicted | All outcomes as reported by the authors |
| **6.Time span of prediction** | Define time period within which the outcome is predicted | All time periods as reported by the authors |
| **7.Intended moment of using the model** | Time point of intended use of the models may be specified | All time points as reported by the authors |

*Note: Adapted from Moons KGM, de Groot JAH, Bouwmeester W et al. Critical Appraisal and Data Extraction for Systematic Reviews of Prediction Modelling Studies: The CHARMS Checklist. PLoS Med. 2014;11(10): e1001744.*

After searching the literature, inclusion and exclusion criteria were then applied as detailed in Chapter 3 to identify eligible articles. In the next section (2.2.2(b)), I present the procedures used to extract data from the eligible articles and the criteria used to critically appraise these data.

*b) Data extraction and critical appraisal*

The second part of the checklist CHARMS provides is a list of items relevant for data extraction. These items are summarised in Table 2.3 while the complete checklist can be found in Appendix 1. For each domain, items for data abstraction are specified alongside recommendations for the evaluation of risk of bias. Data on each of these domains were extracted from each identified article and entered into a CHARMS data extraction template (Appendix 2).

***Table 2.3**: Relevant items for data extraction and critical appraisal.*

| Domain | Example items | Example recommendations | Potential limitations/bias |
|---|---|---|---|
| **1. Source of data** | Design of the study that provided the data set | Prospective cohort preferred | Missing data in retrospective cohorts |
| **2. Participants** | Inclusion and exclusion criteria | Should be representative of target population | Intensive care unit models not applicable to facilities offering essential neonatal care |
| **3. Outcome(s) to be predicted** | Definition and method of measurement | Should match that specified in the review question | All-cause mortality versus disease specific mortality |
| **4. Candidate predictors** | Number, definition and method of measurement | Assessors should be blind to outcome | Inflation of predictive accuracy of the predictors if not blinded |
| **5. Sample size** | Dichotomous outcomes (events) | 10 events per predictor for model development | Inaccurate predictions in external population (overfitting) |
| **6. Missing data** | Frequency of missingness | Multiply impute missing data | Exclusion of missing data reduces representativeness |
| **7. Model development** | Multivariable regression modelling | Inclusion of all candidate predictors (full model) | Predictor selection procedures may lead to selection bias |
| **8. Model performance** | Model accuracy | Discrimination and calibration | Classification measures e.g. sensitivity and specificity biased |
| **9. Model evaluation** | External validation | At least 100 events and 100 non-events in the population | Random split of one dataset increases risk of bias |
| **10. Results** | Final model formula | Regression equation | Not applicable |
| **11. Interpretation and Discussion** | Strengths, weaknesses, challenges | Use PRISMA | Not applicable |

Note: Adapted from Moons KGM, de Groot JAH, Bouwmeester W et al. Critical Appraisal and Data Extraction for Systematic Reviews of Prediction Modelling Studies: The CHARMS Checklist. PLoS Med. 2014;11(10): e1001744.

Each item of data extracted was assessed for the presence of limitations or bias using the recommendations as exemplified in column three of Table 2.3.

The Cochrane Bias Methods Group defines bias as “a systematic error, or deviation from the truth, in results or inferences”.(109) In systematic reviews, bias arises from the characteristics of the primary studies included. The CHARMS checklist identifies limitations and bias in *individual* articles (Table 2.3). However, CHARMS

does not quantify the extent to which the potential bias may have affected the results *across all* the identified articles. To tackle this, the **P**rediction study **R**isk **O**f **B**ias **A**ssessment **T**ool (PROBAST) is under development for the assessment of risk of bias and applicability of prediction modelling studies.(110) As PROBAST was not available at the time of conducting this systematic review, the **G**rading of **R**ecommendations **A**ssessment, **D**evelopment, and **E**valuation (GRADE) system was used as a guiding framework for this part of the review (2.2.3 ).(111)

### *2.2.3 Grading of recommendations assessment, development, and evaluation (GRADE)*

The GRADE system was developed by the GRADE working group to address the confusion that resulted from having multiple systems of rating quality of evidence.(111) The parts of the GRADE process are outlined in Box 2.1.

1. Framing the research question.
2. Specification of outcomes and rating their importance.
3. Evaluation of the evidence.
4. Recommendations taking into consideration the values and preferences of patients and society.

***Box 2.1:*** *The Grading of recommendations assessment, development, and evaluation (GRADE) process.*
Adapted from Guyatt G, Oxman AD, Akl EA, Kunz R, Vist G, Brozek J, et al.
GRADE guidelines: 1. Introduction—GRADE evidence profiles and summary of findings tables. Journal of Clinical Epidemiology. 2011;64(4):383-94.

The research question and outcomes were defined using CHARMS guidance (Section 2.2.2 (a)). The GRADE method was then used to (1) rate the importance of the outcomes (section 2.2.3 (a)) and (2) to evaluate of the quality of the evidence across all identified articles after (Section 2.2.3(b)). I thereafter extracted data from the individual studies and conducted critical appraisal. Data extraction and appraisal was also done independently by a second reviewer (Michuki Maina) with disagreements resolved by discussion. (112)

*a) Rating the importance of outcomes using GRADE*

GRADE provides two methods for rating the relative importance of outcomes; three categories or the equivalent on a 9-point scale as shown in Table 2.4.(113, 114)

***Table 2.4:*** *Ranking the importance of outcomes using GRADE.*

| Category | Point scale |
|---|---|
| Critical for decision making | 9,8,7 |
| Important but not critical for decision making | 6,5,4 |
| Not important for decision making | 3,2,1 |

Note: Adapted from Guyatt GH, Oxman AD, Kunz R et al. What is "quality of evidence" and why is it important to clinicians? BMJ. 2008;336(7651):995-8 and Guyatt GH, Oxman AD, Kunz R et al. GRADE guidelines: 2. Framing the question and deciding on important outcomes. Journal of Clinical Epidemiology. 2011;64(4):395-400

The GRADE approach helps to focus attention on outcomes that are most important in decision making regarding rating the quality of evidence and later in making recommendations for clinical policy.(114) In addition, after retrieval of the relevant studies, new outcomes identified in the articles but not initially considered can be rated in the same way.(114) In this systematic review I began by specifying general outcomes in keeping with the intention to have a broad search strategy. These were, mortality, morbidity and resource utilisation. Outcomes reported in the eligible articles were classified into these three groups and the importance of each outcome across all articles was then rated using the three categories in Table 2.4.

*b) Rating the certainty of the evidence using GRADE*

Certainty of the evidence (formerly called quality of the evidence) is the extent to which we have confidence in the estimates reported in the primary studies or alternatively the "extent to which the effect estimates (study results) are sufficiently credible to support a particular recommendation".(115, 116) In GRADE this is assessed by five criteria (Table 2.5) that may lower certainty of the evidence pertaining to each of the outcomes of interest separately.(113) There are also

criteria for upgrading the certainty in the evidence but this approach was not used in this review for reasons given in Table 2.6.(117)

***Table 2.5:*** *Evaluation of certainty in the evidence across studies using GRADE.*

| Criteria | Description |
|---|---|
| **1. Study limitations** | Limitations in study design that may lead to biased results (118) |
| **2. Inconsistency of results** | Differences between studies in the outcomes of interest (119) |
| **3. Indirectness of evidence** | Populations in identified studies different from population of interest (120) |
| **4. Imprecision** | Wide confidence intervals in the study results (121) |
| **5. Reporting bias** | Extent to which probability of publication is influenced by the results (122) |

*Potential reasons to lower certainty in the quality of the evidence
Note: Adapted from Guyatt GH, Oxman AD, Kunz R et al. What is "quality of evidence" and why is it important to clinicians? BMJ. 2008;336(7651):995-8

There is no specific GRADE guidance for prediction modelling studies. The GRADE guidance for diagnostic tests was used as a framework for this purpose because of similarities with the CHARMS checklist.(123) An important similarity is that the development of diagnostic tests typically uses data from observational studies. I therefore used the supplementary table from Gopalakrishna et al as a guide for GRADE criteria to assess certainty of the evidence for prediction modeling studies.(124) The original table by Gopalakrishna and colleagues is included in Appendix 3. These GRADE criteria presented in Table 2.6 are thus novel aspects to this DPhil.

***Table 2.6:*** *Application of GRADE to assess certainty of evidence in prediction modelling studies.*

| Category | GRADE for Prediction Modelling Studies |
|---|---|
| **1.Question formulation** | PICOT formulation; Population Issue Context Outcome Time<br>Used CHARMS to guide important issues to consider* |
| **2.Outcomes** | Patient important outcomes, CHARMS guidance |
| **3.Rating the quality of evidence** | By outcome, across studies |
| **4.GRADE criteria for downgrading** | |
| Risk of bias (RoB) | 1. First, assess RoB for each study. Use RoB domains in CHARMS<br>2. Downgrading is based on considered judgment on the extent of bias in all studies reporting the outcome |
| Indirectness | Downgrading based on applicability issues. Use applicability domains of CHARMS tool. |
| Inconsistency | Downgrading based on evaluation of unexplained heterogeneity.<br><br>Criteria are less clear compared to intervention studies, but include judgment on similarity of point estimates and extent of overlap of confidence intervals. Lack of appropriate statistical methods for assessing study heterogeneity. |
| Imprecision | Downgrading is based on evaluation of 95% confidence intervals around model discrimination and calibration. Formal guidance and methods on how this can be assessed not available yet. |
| Publication bias | Judgment of this criterion is challenging.<br><br>Existing methods used in intervention studies not applicable to prediction modelling studies |
| **5.GRADE criteria for upgrading** | Specific examples of situations in which upgrading is appropriate not yet available. |

*Table 2.2 page 37

Note: Adapted from Gopalakrishna G, Mustafa RA, Davenport C Et al. Applying Grading of Recommendations Assessment, Development and Evaluation (GRADE) to diagnostic tests was challenging but doable. Journal of Clinical Epidemiology. 2014;67(7):760-8.

For each outcome, the second reviewer and I started off with high certainty as the default rating since as noted in category 5 of Table 2.6, Gopalakrishna and colleagues found no examples in which upgrading is appropriate. We then applied these GRADE criteria to consider reasons for downgrading (Table 2.6, category 4). Following this assessment, the certainty of evidence for each outcome was rated

using a four-level scale illustrated in Table 2.7. This table is also novel to this DPhil as I modified the terms used (for diagnostic tests) to reflect its intended use in the evaluation of prediction models.

***Table 2.7:*** *Rating of certainty of evidence in predictive model predictive performance*

| Rating | Interpretation |
|---|---|
| **1. High certainty** | Further research is very unlikely to change our confidence in the estimate of predictive performance |
| **2. Moderate certainty** | Further research is likely to have an important impact on our confidence in the estimate of predictive performance and may change the estimate |
| **3. Low certainty** | Further research is very likely to have an important impact on our confidence in the estimate of predictive performance and is likely to change the estimate |
| **4. Very low certainty** | Any estimate of predictive performance is very uncertain |

Note: Adapted from Guyatt GH, Oxman AD, Vist GE, Kunz R et al. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. BMJ. 2008;336(7650):924-6.

## 2.3 Study site and data collection procedures

The previous section (Section 2.2) provides an overview of methods used in the systematic review of neonatal treatment intensity scores (Objective 2(a)). In this section (2.3) and in the subsequent sections (2.4 to 2.7), I now describe the methods used in the primary research conducted to answer Objectives 1,2(b),3 and 4.

### *2.3.1 Study site*

Data were collected in the largest maternity hospital in Nairobi County. This facility provides primary level antenatal, intrapartum and postpartum care but also serves as a first referral level facility (receives patients from primary care facilities rather than referrals from other hospitals). In the period covered by the study data, the hospital had approximately ≈20,000 deliveries and ≈5000 admissions to the neonatal unit annually. In a recent survey (September 2015) the hospital had in total an estimated 130 nurses, two clinical officers, 25 medical officers, three paediatricians and five obstetricians. (Personal communication, study results yet to be published).(82)

The neonatal unit is run by a clinical team consisting of general paediatricians, medical officers (non-specialist doctors), nurses (no specialist neonatal nurses) and clinical officers (physician assistants). On a typical day in the unit, the staff on duty are as illustrated in Table 2.8 (personal communication, study results yet to be published).(82)

***Table 2.8:*** *Staff on duty in the newborn unit on a typical shift in the study hospital*

| Cadre | Weekday | | Weekend | |
|---|---|---|---|---|
| | Day | Night | Day | Night |
| **Student nurses** | 10 | 1 | 4 | 1 |
| **Nurses** | 6 | 4 | 4 | 3 |
| **Clinical officers** | 2 | 1 | 1 | 1 |
| **Medical officers** | 2 | 0 | 1 | 0 |

Note: these numbers do not include on-call staff and the 3 general paediatricians. Data from personal communication, study results yet to be published.(82)

This unit principally offers essential neonatal care as defined in Chapter 1 (this includes radiant warmers, phototherapy, cots and incubators, intravenous fluids, enteral feeding, antibiotics, kangaroo mother care and supplemental oxygen). There are no facilities for more advanced or critical care except two BCPAP machines acquired during the latter part of data collection. While the high workload is not typical of a neonatal unit in Kenyan public hospitals, the level of care available is (except for BCPAP).

### *2.3.2 Data collection procedures*

The data sets used for survival analysis (Objective 1) and clinical prediction modelling (Objectives 2b,3 and 4) were obtained from a routine data collection system set up in the neonatal unit of the study hospital. I led the development and implementation of this system as part of a collaboration between the Health Services Unit of KEMRI-Wellcome Trust, the National Ministry of Health, Nairobi County Government, Kenya Paediatric Association and the hospital's administration. Data collection commenced in February 2014 and continues to date. The overarching aim is to improve the availability and quality of neonatal data to augment the existing hospital information system. These data are primarily used by the neonatal clinical team to understand and improve the quality of care.

A trained data clerk based in the neonatal unit enters data for all patients at the point of exit from the unit. The source documents are: the standardised admission proforma (neonatal admission record; NAR), treatment sheets, fluid monitoring charts, feed charts, vital sign charts, diagnostic tests, nursing notes and the discharge summary. The focus is on clinical information documented at admission in the NAR and at the point of discharge. The key variables are: demographic data, symptoms and signs, diagnoses, diagnostic tests, treatments prescribed and the outcome at the point of exit from the unit.

Data are abstracted from the patient file and entered directly using REDCap™ electronic data capture tools. REDCap™ (**R**esearch **E**lectronic **D**ata **Cap**ture) is a secure, web-based application designed to support data capture for research studies, providing (1) an intuitive interface for validated data entry; (2) audit trails for tracking data manipulation and export procedures; (3) automated export procedures for seamless data downloads to common statistical packages; and (4) procedures for importing data from external sources.(125) Given the large number of patients in the unit, data is entered on a select number of variables for all patients (minimum data set) while data on all the variables (full data set) is entered on 60% of patients. At the point of data entry, the REDCap™ system randomly assigns each new record to either minimum or full data set. The 60% sample is based on the estimated number of records that a data entry clerk can comfortably enter in a day as observed in previous surveys my colleagues and I have undertaken in the Health Services Unit of the KEMRI-Wellcome Trust Research Programme.(32, 57, 126, 127)

*2.3.3 Data management*

Data entry is guided by a standard operating procedure. Data are examined for errors in real time using checks built in REDCap™ and at the end of each day using an R algorithm (R script). R is an open source programming language that provides an environment for statistical computing and graphics.(128) Any errors identified are corrected, where possible, by referring back to the source document. The clean data file is then stored locally (in the neonatal unit) in the data entry computer and a copy is uploaded onto a central server in KEMRI-Wellcome Trust. A final error check is then done using an R script leading to a final round of corrections where possible. A feedback report is generated every three months using these data and is used by the clinical team to identify areas that require improvement of documentation in the clinical notes (paper records). This process is illustrated in Figure 2.1.

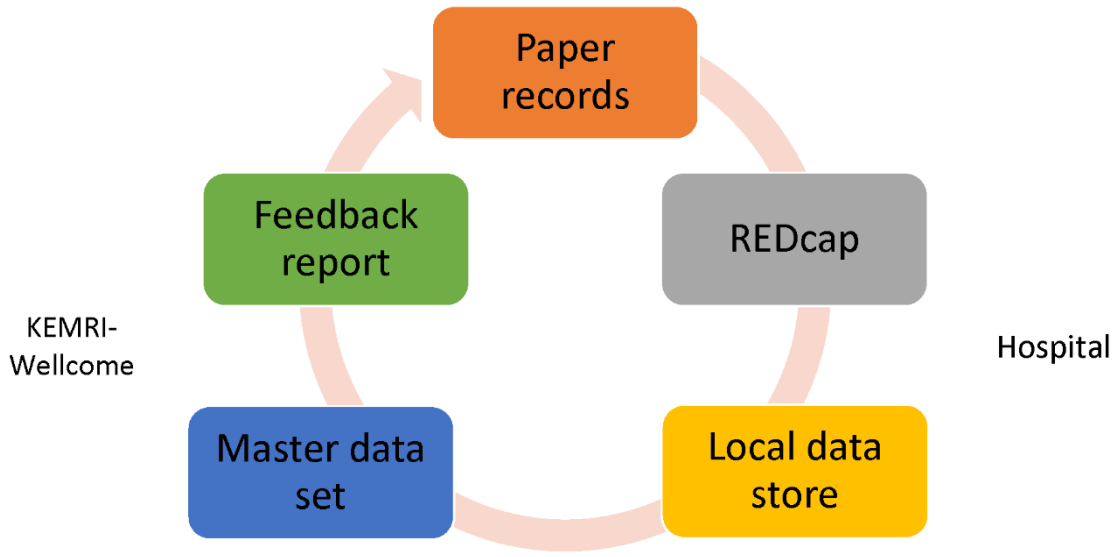



***Figure 2.1**: Data management processes*

Further data quality assurance is done every three months by randomly sampling existing records and retrieving the patient file for double data entry to analyse for concordance. Access to the data sets is password protected both at the hospital and at KEMRI-Wellcome Trust. Ethical considerations are described in Section 2.9.

## 2.4 Description of the study data sets

From the data collected, two data sets were extracted for this study (Table 2.9). The first data set includes admissions from April 2014 to December 2015 (n=9115) while the second includes admissions from January 2016 to July 2016 (n=2735). In each data set, 60% of the observations had data on all the variables entered (full data set) and 40% contain the restricted set of variables (minimum data set). A list of the variables included in this routine data collection system (Section 2.3.2) is provided in Appendix 4.

***Table 2.9:*** *Description of the study data sets*

| Data set | Period | Number of observations | |
|---|---|---|---|
| | | Total* | Random sample† |
| **Data set 1** | April 2014 – December 2015 | 9115 | 5427 |
| **Data set 2** | January 2016 – July 2016 | 2735 | 1627 |

* Contains restricted set of variables (minimum data set)
† 60% of the total randomly selected at the point of data entry to include of the full set of variables

Data set 1 (n=9115) was used to describe the clinical characteristics of the study population and to examine the pattern of survival after admission using statistical methods outlined in Section 2.5 (Objective 1). A subset of Data set 1 (n=5427) was utilised to derive two novel prognostic models for study Objectives 2 and 3 as described in Section 2.6. External validation of both models - as described in Section 2.6 - was subsequently done using a subset of Data set 2 (n=1627). Detailed descriptions of these data sets and the variables used in the analyses are provided in Chapter 4 for survival analysis and Chapter 5 and 6 for the prognostic models.

## 2.5 Survival analysis

### *2.5.1 Introduction*

The first objective of this study was to describe the clinical characteristics and analyse in-hospital survival of neonates admitted to the neonatal unit of the study hospital. The main component of this objective is the description of the time to death in hospital after admission.

Approaches to time to event analysis include:

i. Non-parametric analysis e.g. Kaplan-Meir and Life table analysis (129)
ii. Semi-parametric analysis e.g. Cox regression (130)
iii. Parametric analysis e.g. Weibull and Gompertz models (130)

I first used Kaplan-Meir survival analysis to describe time to in-patient death after admission and thereafter extended this to Competing Risk survival analysis. These are described in Section 2.5.2 and 2.5.3 respectively.

### *2.5.2 Kaplan-Meier survival analysis*

In the data set used, the follow up period for survival analysis commenced on the date of admission and ended on the date of exit from the neonatal unit. There are three possible outcomes at the point of exit from the neonatal unit: death, discharged alive or referred to another heath facility. Using the Kaplan-Meier method the probability of being alive and still in the neonatal unit was computed at each time point an in-hospital death was observed. This was done in R: A language and environment for statistical computing using the “Survival” package.(128, 131)

Patients who were either discharged alive or referred were censored (right censored) in these analyses.

The key assumption underlying the right censoring is that death in the study neonatal unit would eventually be observed in those who were censored if they were followed up. However, this assumption is violated as the outcome of interest is in-hospital death after primary admission in the neonatal period. In addition, the study neonatal unit does not re-admit patients after discharge or referral. For this reason, the competing risk survival analysis method was used to complement the results of the Kaplan-Meier analysis. A detailed description of the Kaplan-Meier analyses is provided in Chapter 4.

*2.5.3 Competing Risks survival analysis*

The Kaplan-Meier method assumes that the outcome of interest will ultimately occur in the censored subjects in the long run.(132) However, the occurrence of an event may preclude the subsequent observation of another. Such events are described as Competing Risks as their occurrence is mutually exclusive.(132, 133) In this study, if a patient was discharged alive or referred (right censored) it meant that in-hospital death (outcome of interest) could not be observed subsequently. This competing risk situation is illustrated in Figure 2.2.

**The image originally presented here cannot be made freely available via ORA because of copyright. The image was sourced from** *Varadhan R, Weiss CO, Segal JB et al. Evaluating Health Outcomes in the Presence of Competing Risks: A Review of Statistical Methods and Clinical Applications. Medical Care. 2010;48(6): S96-S105*

***Figure 2.2:*** *In-hospital death and discharge alive/referral as competing risks. Adapted from Varadhan R, Weiss CO, Segal JB et al. Evaluating Health Outcomes in the Presence of Competing Risks: A Review of Statistical Methods and Clinical Applications. Medical Care. 2010;48(6): S96-S105*

Kaplan-Meier analysis in the presence of competing risks leads to overestimation of probability of the event of interest.(132, 133) The competing risk method overcomes this limitation by simultaneously estimating the probability of the outcome of interest and the probability of the competing event.(132) We therefore used this approach to compute the probability of in-hospital mortality and probability of being discharged alive.

The computation of these competing risks was done using the cumulative incidence function method. The probability of any event is partitioned into the probability of the outcome of interest and the probability of the competing risk.(132) In this study, the probability of all outcomes is given by the sum of the probability of in-hospital death (outcome of interest) and the probability of discharge alive (competing risk). The referrals were censored. These analyses were implemented in R using the "cuminc" function of the "cmprsk" package.(134) A detailed description is provided in Chapter 4.

## 2.6 Prognostic modelling

### *2.6.1 Introduction*

In Chapter 1, I outlined the four phases in the development of prediction models: development, validation, impact assessment and implementation. I developed and validated two prognostic models using six out of the seven steps recommended by Steyerberg and colleagues.(135) These six steps are summarized in Table 2.10. The seventh step deals with clinical utility of the model and I describe this separately in Section 2.7.

***Table 2.10**: Summary of methods used in prediction model development and validation*

| Step | Description |
|---|---|
| **Development** | |
| 1. Problem definition and data inspection | Definition of endpoint of interest and description of data sets used |
| 2. Coding of predictors | Predictors specified as either continuous or categorical |
| 3. Model specification | Selection of statistical modelling approach |
| 4. Model estimation | Procedures during modelling |
| 5. Model performance | Quantification of predictive accuracy |
| **Validation** | |
| 6. Internal and external validation | Estimation of overfitting (internal validation) and performance in an external population (external validation) |

Adapted from Steyerberg EW, Vergouwe Y. Towards better clinical prediction models: Seven steps for development and an ABCD for validation. European Heart Journal. 2014;35(29):1925-31.

The two models developed are named: Neonatal Essential Treatment Score (NETS) and Score for Essential Neonatal Symptoms and Signs (SENSS). The outcome of interest for both models was in-hospital death after admission to the neonatal unit. Two data sets were used for model development and validation as described in Section 2.4. Predictors were treatments prescribed at admission for NETS and symptoms and signs of severe illness at admission for SENSS.

In addition, I included birthweight and sex in both models. The specifics of problem definition (including justification for predictor selection), data inspection and coding of predictors are provided in Chapter 5 (NETS) and Chapter 6 (SENSS). I describe the general aspects of statistical approaches to model specification, estimation, performance and validation in the next sections (2.6.2 to 2.6.5). The specific details of the application of these methods are provided in Chapters 5 and Chapter 6.

### 2.6.2 *Model development*

*a) Approaches to model development*

Dorling and colleagues have described two ways of developing prediction models: "Medical" scores developed by experts through consensus and "statistical" scores which are derived from data.(102) Statistical methods may be further classified into two: machine learning and regression modelling. In this study I used regression modelling, specifically logistic regression. I used this method because the outcome of interest is binary (dead or not dead at exit) and the time to event of interest is short (maximum follow up 98 days).

*b) Logistic regression in prediction modelling*

The probability of in-hospital death was predicted by a linear combination of variables (predictors) and their regression coefficients (odds ratios on the log scale).(136) These coefficients were estimated by the maximum likelihood approach. First the linear predictors (i.e. regression coefficients multiplied by the predictor values for each individual in the data set) were estimated as illustrated in Box 2.2. Thereafter, the linear predictors were converted to probability of death on a [0,1] interval using the formula in Box 2.3.(136)

**Linear predictor (lp)= $\alpha + \beta_i * x_i + \ldots B_j * x_j + \varepsilon$**

Where α = baseline odds of death when none of the predictors are present

β = the coefficients, in this case odds ratios on the natural log scale

X = the predictors

ε = error term (the difference between observed and predicted)

**Box 2.2:** Estimation of the linear predictor of death in logistic regression

**Probability of death= $e^{lp} / (1 + e^{lp})$**

Where $e^{lp}$ =the natural log exponent of the log odds of death

***Box 2.3:*** *Calculation of predicted probability of death from odds of death*

These logistic regression models were implemented in R using the “lrm” function of the “rms” package.(128, 137) Further details on model specification and estimation are provided in Chapter 5 (NETS) and Chapter 6 (SENSS).

c) Missing data

Missing data are a common phenomenon in research.(138-140) In the context of prediction research this refers to the presence of missing predictor and outcome values. Data from routine care such as those used in this study may be particularly prone to missingness and as such this was a key consideration during model development and validation.(101)

There are three mechanisms that may lead to missing data: missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR).(141) Under MCAR, the probability of missing is completely unrelated to observed (variables included in the study) and unobserved (unknown) patient characteristics meaning individuals with missing data are a random sample representative of the population with complete data.(136, 142) In MAR, the probability of missing is dependent on observed values.(136) MNAR means that the probability of missing data depends on either the values that are missing or on unobserved characteristics.(136)

I used logistic regression to develop the two prediction models for this study. In regression analysis, statistical software typically deletes any individuals with missing values. This is termed “complete case analysis”.(136, 139) Such an approach leads to two key problems. Firstly, any missingness leads to deletion of

all the data on the individual including available data leading to a reduction in sample size and creating statistical inefficiency as information is discarded.(136) Secondly, this approach may lead to bias in the estimated regression coefficients if there are systematic differences between the individuals with completed data and those with missing data because complete case analysis makes the assumption of MCAR.(136, 139)

In handling the missing values in the data sets, I considered three key questions: (1) Why are data missing?, (2) How do patients with missing and complete data differ? and (3) Do the observed data help predict the missing values? (143) The preferred approach to handling missing data in prediction is multiple imputation.(101, 144) In this approach, the available data is used to estimate plausible values for the missing data using regression analysis.(101, 136) Detailed description of patterns of missingness of specific variables and how these were handled including multiple imputation procedures where applicable are provided in Chapter 5 and Chapter 6.

d) Handling of continuous predictors

Continuous predictors are quantitative variables that can take any value within a given range. In generalised linear models such as logistic regression which was used in this study, we assume that the effect of the continuous predictor on the outcome is linear i.e. its effect on outcome is the same throughout the range of predictor values.(136) This linearity assumption is tested by fitting non-linear relationships between the continuous predictor and the outcome and assessing if

this improves the model fit.(136) There are various methods of examining these non-linear effects including polynomials, fractional polynomials and splines.(136)

Alternatively, the continuous predictor may be categorised. The advantages of this approach include that it simplifies the analysis and makes it easier for clinicians to use the model.(144) However, categorisation is generally not recommended as it results in loss of information about predictor effects particularly when two categories are used (dichotomisation).(144) Categorisation into more groups reduces the loss of information and is common in epidemiological studies.(136, 144)

In this study there is only one continuous predictor used i.e. birthweight. Birthweight is however included as a categorical predictor in the two models developed for three reasons. First, the categories used are firmly established in clinical practice and therefore clinicians are well versed with them. Secondly in routine practice five weight categories are used meaning less loss of information in contrast to two categories. Lastly there is strong evidence that low birth weight is associated with risk of death which is the outcome of interest in this study.(52)

### *2.6.3 Model performance*

The evaluation of model performance is defined as the quantification of how good the predictions from the model are.(136) There are two key characteristics for measuring predictive performance, discrimination and calibration.

a) Discrimination

Discrimination is the ability of the model to distinguish between an individual with the outcome of interest and an individual without. In this study, this is the models' ability to distinguish between patients who died in hospital from those who were alive at the point of exit. Discriminative ability was measured using the concordance statistic (c-statistic). For binary outcomes the c-statistic is equivalent to the area under the receiver operator curve.(135, 136) This estimates the probability that for all possible pairings of subjects in the data set, those who died have a higher predicted risk of death compared to those who were discharged alive. A value of 0.5 means that the model is no better than flipping a coin (no better than chance) whilst a value of 1 is perfect discrimination.(136)

b) Calibration

Calibration refers to the agreement between observed outcomes and predictions. In this study, it is the extent to which predicted deaths match the observed deaths in the data. This was estimated using calibration curves implemented in the R package "CalibrationCurves".(135, 136, 145) The calibration curves were constructed by plotting the observed deaths on the *y*-axis against the predicted probability of death on the *x*-axis. Since observed death is binary (i.e. 1 if dead, 0 if alive) and the predicted probability of death is on a continuous scale from 0 to 1,

the predicted probabilities were grouped into tenths of predicted probability before plotting. For each tenth of predicted probability, the proportion with observed death was plotted on the *y*-axis against the predicted probability on the *x*-axis.(136)

Two measures were then calculated: the calibration-in-the-large and the calibration slope. (135, 136) Calibration-in-the-large is the intercept of the calibration curve and it is the difference between the mean observed and the mean predicted risk of death. In a perfectly calibrated model, this is equal to zero (i.e. the difference is zero). The calibration slope was estimated using logistic regression where the outcomes (deaths) were regressed on their predicted risk. In a perfectly calibrated model the slope is equal to 1 but it is much less than 1 in a model with overfitting.(146) Overfitting is described in section 2.6.4.

### *2.6.4 Internal validation*

Internal validation evaluates model performance in patients from the same underlying population; the reproducibility of the model.(147) The regression modelling process as described in section 2.6.2 finds the best fit of the chosen predictors in estimating the probability of the observed outcomes. This, as noted by Collins and Moons, usually results in models performing better on the dataset from which they were developed in comparison to existing models developed in external populations; a phenomenon referred to as *overfitting*.(136, 148) Model performance values can therefore be optimistic as a result of the modelling process and overfitting.

Consequently, we need a method to quantify and correct for model optimism. I used the bootstrapping method to quantify the optimism of the models in the development data set using the procedures in Box 2.4.(136) The bootstrapping approach was implemented using the "validate" function of the R package "rms".(137)

1. Model specification and estimation in the development data set (original model). Determine the apparent performance.
2. Draw a bootstrap sample with replacement from the development data set.
3. Specify and estimate the model in the bootstrap sample (bootstrap model). Determine bootstrap sample performance (bootstrap performance).
4. Apply the bootstrap model to the development data set. Determine performance (test performance)
5. Calculate optimism: Optimism=Bootstrap performance – Test performance
6. Repeat steps 1-4 1000 times to obtain a stable estimate of the optimism
7. Subtract the final optimism estimate from the apparent performance to obtain the optimism-corrected performance estimate

***Box 2.4**: Calculation of optimism-corrected performance for prediction models.*
Adapted from Steyerberg EW. Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating. New York: Springer; 2009

### *2.6.5 External validation*

External validation refers to the evaluation of model performance in a population external to that used to develop it; the transportability of the model.(106, 136, 147) Deterioration in model performance (with respect to discrimination and calibration) in external populations may occur due to *underfitting*.(147) According to Justice and colleagues, underfitting occurs when important predictors of the outcome are omitted from the model (missed predictors).(147) This phenomenon is related to the principal determinant of deterioration in performance; the case-mix. In prediction modelling, case-mix refers to the distribution of predictors that are included in the model as well as the missed predictors.(136) Differences in case-mix may arise due to differences in study settings (e.g. model developed in

neonatal intensive care and applied to a setting offering essential neonatal care only) or time.(136)

I externally validated the two models by measuring predictive performance in a separate data set comprising patients who were more recently admitted to the study hospital. This form of external validation is referred to as temporal, historical or narrow validation as the model was tested in the same hospital setting but on patients from a different time point.(136, 144) Other approaches include geographical or broad validation (testing in different sites) and fully independent validation by different investigators in different sites.(136)

Predicted outcomes in the external validation data set were first obtained using the built-in R function “predict.lm”. Predictive performance (model discrimination and calibration) was thereafter assessed using calibration curves implemented in the R package “CalibrationCurves”.(145)

## 2.7 Clinical utility of prediction models

Model discrimination and calibration which were described in Section 2.6.3 are statistical measures of performance which do not provide guidance on how the models should be used in clinical practice. A key aspect of clinical decision making is evaluation of the balance between the benefits and harms for different decisions. In the case of prediction models this means determining whether basing clinical decisions (e.g. interventions) on a model would do more harm than good.(149) This would entail selecting a cut-off for the range of predicted probabilities to classify patients as positive (above cutoff) or negative(below cutoff).(136) A cutoff for a prediction model is referred to as a “decision threshold” or “classification cutoff”.(136)

Decision analysis provides a means to determine prediction model cut-offs by mathematically assessing the balance between benefits and harms.(136, 150) This yields an optimal decision threshold. In the case of treatment decision making, individuals with predicted probability above the threshold should receive the intervention in question whilst those below should not.(136) Typical decision analysis however requires extra data that are not found in the validation data set on benefits e.g. number of life-years saved and costs e.g. admission costs.(136, 150) In addition decision analysis would require that the prediction model gives a binary result while predictions are typically in the range of 0-1 (or risk 0-100%).(150) To overcome these challenges, a simplified approach to decision analysis has been developed, “net benefit”.(149, 150)

I therefore used this net benefit approach to assess the potential clinical utility of the two models I developed (NETS and SENSS). In this method, decision curves are constructed by plotting the net benefit on the *y*-axis against the range of predicted probabilities on the *x*-axis.(149) I computed the net benefit across the range of predicted probabilities of death (0-1) for each model using the procedures in Box 2.5 below as outlined by Vickers and colleagues.(149) These computations were done on the external validation data set.

1. All the predicted probabilities of death were considered as a possible threshold probability (***Pt***) to define when a patient is positive (i.e. in-hospital death observed)
2. For each threshold probability(***Pt***), patients with predicted risk ≥ ***Pt*** who died were counted as true positive while those who did not die were counted as false positives
3. With N the total sample size, the net benefit was calculated using the equation:

$$\textbf{\textit{Net benefit}} = \frac{\textbf{\textit{True positives}}}{\textbf{\textit{N}}} - \frac{\textbf{\textit{False positives}}}{\textbf{\textit{N}}} \times \frac{\textbf{\textit{Pt}}}{\mathbf{1 - \textit{Pt}}}$$

4. Steps 2 and 3 were repeated from the least to the highest predicted probability of death for each model.
5. Steps 1-4 were repeated for 2 assumptions: all patients would experience in-hospital death and no patients would experience in-hospital death.

***Box 2.5****: Calculation of net benefit across the range of threshold probabilities.*
Adapted from Vickers AJ, Van Calster B, Steyerberg EW. Net benefit approaches to the evaluation of prediction models, molecular markers, and diagnostic tests. BMJ. 2016;352.

This computation of the net benefit considers the performance of the prediction model (detection of true positives and false positives) across the range of predicted probabilities. To illustrate this let us consider the prediction of mortality where we have an intervention that benefits only those likely to die (true positive) but has no value to those who are not likely to die (false positive). If we consider each predicted probability as a potential cutoff, those classified as positive and therefore

eligible to receive the intervention will include true positives plus the false positives. Subtracting the false positives from the true positives leaves only those who would truly benefit from the intervention which is thus referred to as the “net benefit”. The unit of net benefit is true positives.

If we for example choose to intervene for everyone with a predicted probability of 0.1 and above, it means that for every ten patients at risk only one will experience the outcome. The implication is that for every ten patients receiving the intervention, one would benefit (true positive) and nine would not (false positive). Going back to the net benefit formula, this ratio of 1:9 provides the weight which is multiplied by the false positives. A hypothetical example is given in Box 2.6.

1.In a hypothetical sample of 1200 patients (N), using predicted probability of outcome of 0.1 as a cut-off (pt) gives 200 true positives (TP) and 400 false positives (FP)

2. Net benefit is given by:

$$\textbf{Net benefit} = \frac{\textbf{True positives}}{\textbf{N}} - \frac{\textbf{False positives}}{\textbf{N}} \times \frac{\textbf{Pt}}{\textbf{1} - \textbf{Pt}}$$

3.In this case net benefit is:

(200/1200) -(400/1200) * (0.1/1-0.1) =0.17

4.Net benefit as a percentage

0.17*100=17%

***Box 2.6****: Illustration of calculation of net benefit at a given cut-off*

Further discussion on the interpretation of net benefit is provided in Chapter 7.

Decision curves were thereafter produced by plotting the net benefit on the *y* axis against the range of predicted probabilities on the *x* axis. Specific details on these analyses comparing the net benefit of the NETS model against the SENSS model are provided in Chapter 7. These analyses were implemented using the "plot_decision_curve" and "plot_clinical_impact" functions of the R package "DecisionCurve".(151)

### 2.8 Reporting of prediction modelling studies

The Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) statement was used to report the two models in Chapter 5 and Chapter 6. (92) The TRIPOD statement is a 22-item checklist that guides the reporting of the design, conduct, analysis and interpretation of prediction modelling studies.

### 2.9 Ethical considerations

This study comprises secondary analyses of previously collected anonymised data. These data were collected in a technical collaboration with the Ministry of Health, Government of Kenya for which ethical approval was obtained from the KEMRI Ethics Review Committee (SSC. Protocol 2771). The original approval and subsequent annual renewals are included in Appendix 5. As this was a secondary analysis of previously collected anonymized data, a waiver of ethical approval was obtained from the Oxford Tropical Research Ethics Committee (OxTREC) and is included in Appendix 5.

## Chapter 3 : Systematic review of neonatal treatment intensity scores and their potential application for low resource settings[1]

### 3.1 Introduction

*3.1.1 Role of clinical prediction models in decision making for essential neonatal care*

In Chapter 1 we saw that clinical prediction models combine multiple patient characteristics to calculate the probability of specific end-points (for example in-hospital mortality) for individual patients, thus integrating information to aid decision making.(70, 95, 152) This is illustrated in Figure 1.6. Predicted risk of in-hospital mortality could for instance provide: (1) a better measure of severity of illness for individual patients and (2) better means of comparing neonatal case-mix of different hospitals therefore potentially leading to better treatment decisions and comparisons of hospital performance respectively (Table 1.9).

As outlined in Chapter 1 (Section 1.2.3), high quality of care is necessary if LMICs are to realise the projected substantial reduction in neonatal mortality expected from the delivery of essential interventions at scale in facilities. (4, 5) By facilitating better decision making, prediction models could contribute to the improvement of the quality of hospital care for neonates in LMICs ultimately improving neonatal survival. Given the high burden of neonatal deaths in LMICs, we therefore sought to identify models that predict neonatal in-hospital mortality and are potentially suitable for use in LMICs in general and in Kenya. To this end, I first did a scoping review of neonatal prediction models (Section 3.1.2 (b)). This preliminary step led

[1] Aluvaala J, Collins GS, Maina M, Berkley JA, English M. A systematic review of neonatal treatment intensity scores and their potential application in low-resource setting hospitals for predicting mortality, morbidity and estimating resource use. Systematic reviews. 2017;6(1):248.

to the selection of models that employ a treatment intensity scoring approach as potentially suitable (Section 3.1.2 (c)). The rest of the chapter is then dedicated to the second step; a systematic review to comprehensively evaluate these neonatal treatment intensity scores.

### *3.1.2 Scoping review of prediction models in neonatal medicine*

#### *a) Rationale for evaluating existing prediction models*

In considering the use of prediction models to support decision making in hospital based essential neonatal care in LMICs, we may either develop a new model or choose from among existing models. Collins and colleagues recommend the latter approach before developing new models to avoid waste of resources as described in Chapter 2. (106, 107) Selecting from amongst existing models should however be guided by the evaluation of existing models for performance and suitability to the context.(147)

The case of prediction models in adult cardiovascular medicine underscores the potential value of first evaluating existing models. This field has seen a proliferation of models as illustrated by a systematic review of prediction models for risk of cardiovascular disease in the general population.(105) This review identified articles reporting development of 363 separate models. Out of these, only 132 (36%) were externally validated (i.e. the assessment of model performance in new populations, Chapter 2). This large number of new models of uncertain usefulness is potentially a waste of resources and accordingly the authors' recommend evaluating the usefulness of existing models by tailoring them to suit local settings.(105)

*b) Overview of prediction models in neonatal medicine*

Prediction models may be broadly classified into two categories based on the method of development; "medical" models and statistical models.(102) The patient characteristics included in the models are selected using either clinical expertise ("medical" models) or statistical procedures (statistical models). (153) In this section, I present an overview of models in neonatal medicine under these two categories before narrowing down to statistical models that may potentially be used to predict in-hospital mortality in essential neonatal care settings of LMICs in the following section (3.1.2 (c)).

Neonatal medical models are predominantly diagnostic in that they are used to grade the severity of specific conditions (Table 3.1)

***Table 3.1**: Examples of "medical" models used to grade severity of illness in neonates*

| Model | Purpose | Parameters |
|---|---|---|
| **Apgar**[154] | Assessment of physiological status in the immediate period after birth | Skin colour, heart rate, reflexes, muscle tone, respiration |
| **Silverman-Andersen**[155] | Grading severity of difficulty in breathing | Chest movement, retractions‡ (intercostal, xiphoid), nasal flaring, grunting |
| **Bell**[156, 157] | Grading severity of necrotising enterocolitis* | Systemic, intestinal and radiological signs |
| **Papile**[158] | Grading severity of intraventricular haemorrhage† | Location of bleeding, ventricular dilatation |
| **Sarnat and Sarnat**[159] | Grading severity of central nervous system dysfunction after hypoxia and ischaemia | Consciousness, motor function, pupils, seizures, EEG§ |

* Condition affecting the gastrointestinal tract
† Bleeding into the cavities that hold cerebrospinal fluid in the brain
‡ Movements of the chest indicating labored breathing
§ Electroencephalogram

Perhaps the most well-known of these " medical "models in neonatal medicine is the Apgar score developed by Virginia Apgar in 1953 for the evaluation of the new born infant at 1, 5 and 10 minutes after delivery.(154) Infants are assigned a score ranging from 0 (worst score) to 10 (best score) based on 5 parameters (Table 3.1). Besides its application in assessment of physiological status, in a recent study of more than a million livebirths in the United Kingdom, low Apgar score (0-3) at five minutes after birth was associated with increased risk of early neonatal death (RR 359,95% CI 277 to 466), late neonatal death (RR 31, 95% C.I 18 to 52) and infant death (RR 50,95% C.I 43 to 59).(160)

The Silverman-Andersen score was developed to grade the severity of difficulty breathing (Table 3.1) using a score ranging from 0 (no difficulty breathing) to 10 (maximal difficulty).(155) This score has been adopted in the Kenyan paediatric clinical practice guidelines to guide the use of BCPAP for respiratory support in neonates.(17) Similar to the Silverman-Andersen score, other models have been developed to grade the severity of specific conditions (Table 3.1) but in contrast, these do not assign numerical scores. These include the Bell criteria for necrotizing enterocolitis, Papile classification for intraventricular haemorrhage and Sarnat and Sarnat staging for hypoxic ischaemic encephalopathy.(156-159)

Other "medical" models include the Dubowitz and Ballard scores which use clinical characteristics to estimate gestational age and the Rodwell haematological scoring system for estimating the risk of neonatal sepsis.(161-164) In this work however, the focus is on prognostic models (Figure 1.6,Chapter 1) and therefore these

“medical’ models used for diagnosis or grading the severity of specific conditions are not considered further.

I was most interested in prognostic models-statistical models for predicting a future outcome - specifically those that predict risk of in-hospital mortality (Section 3.1.1). To identify candidate models suited to this purpose, I began by looking at existing reviews of neonatal prognostic models. This was not a systematic review of neonatal prognostic models per se but rather an overview of existing literature as a preliminary step in identifying candidate models for a systematic review (Section 3.1.2 (b)).

To accomplish this, I applied a published search filter for prognostic models validated for PubMed plus MeSH terms for neonates and limited the results to review articles.(165) This search identified 10 review articles; 8 reviewed neonatal models in general while 2 were specific to very low birth weight and very premature neonates.(102, 153, 166-173) A summary of the models potentially applicable to *any* neonate admitted to a neonatal unit as extracted from these articles is presented in Table 3.2. In this table the models are described by name, type of parameters and outcomes predicted. The complete lists of predictors for each model are provided in Appendix 6.

According to Casamassima et al, prognostic models may be classified by the type of parameters used to predict the outcomes.(169) These are:

- Physiological models; based on the extent of derangement of physiological parameters
- Therapeutic models; assume that severity of illness is proportional to number of interventions/treatments required
- Organ-specific models; assume that severity of illness is proportional to the number of organ systems involved.

***Table 3.2**: Summary of prognostic models for predicting in-hospital neonatal mortality*

| Model | Model parameters | | Predicted outcomes | | |
|---|---|---|---|---|---|
| | Physiological | Therapeutic | Mortality | Morbidity | Resource use |
| **NTISS**[104] | | ✓ | ✓ | ✓ | ✓ |
| **SNAP**[174] | ✓ | | ✓ | ✓ | |
| **SNAP-II**[175] | ✓ | | ✓ | | |
| **SNAP-PE**[176] | ✓ | | ✓ | | |
| **SNAP-PE II**[175] | ✓ | | ✓ | | |
| **NMPI**[102] | ✓ | | ✓ | | |
| **Sinkin**[177] | ✓ | | | ✓ | |
| **NBRS**[178] | ✓ | | | ✓ | |
| **TRIPS**[179] | ✓ | | ✓ | | |
| **TRIPS-II**[180] | ✓ | | ✓ | | |
| **VON-RA**[181] | ✓ | | ✓ | | |
| **NQI-2*** | ✓ | | ✓ | | |

*NTISS* (Neonatal Therapeutic Intervention Scoring system), *SNAP* (Score for Neonatal Acute Physiology), *SNAP-PE* (Score for Neonatal Acute Physiology-Perinatal Extension), *NMPI* (Neonatal Mortality Prognosis Index), *NBRS* (Nursery Neurobiologic Risk Score), *TRIPS* (Transport Risk Index of Physiologic Stability), *VON-RA* (Vermont Oxford Network Risk Adjustment), *NQI-2* (Neonatal Quality Indicator)
*Withdrawn, no longer available online

We observe in Table 3.2 that the NTISS stands out as the only therapeutic model.(102) In addition, it is the only model that has been used to estimate resource utilization (especially nursing workload).(104)

The systematic review by Medlock and colleagues identified 41 articles which are not included in Table 3.2 as these are specific to very low birth weight (VLBW) and very premature infants.(173) These include the commonly used Clinical Risk Index for Babies; CRIB.(182) The authors also identified a model developed specifically for use for VLBW babies in LMICs named the Simplified Age-Weight score; a derivative of the CRIB II score.(183) This was developed by Rosenberg and colleagues working in Bangladesh, Nepal and Egypt.(103) Other models for preterm infants include the Berlin(184), National Institute of Child Health and Human Development; NICHHD(185, 186), Australian and New Zealand Neonatal Network; ANZNN (187) Neonatal del Cono Sur Network; NEOCOSUR(188) and Neonatal Multiple Organ Dysfunction; NEOMOD(189) scores. These models were not considered further as I was interested in models that could be potentially applied to all neonates admitted to hospital rather than a specific group such as VLBW or preterm.

*c) Selection of candidate models suitable for predicting neonatal in-hospital mortality in LMICs*

Given the identified limitations in decision making approaches at clinical and service delivery levels in hospital care for neonates in Kenya (Table 1.5), I was interested in models that could provide information to guide such decisions. Risk stratification using predicted in-hospital mortality may be used to guide treatment decisions for individual patients as well as for case-mix adjustment when comparing facilities. Resource use estimation - by providing information to guide strategic and resource allocation decisions - may contribute to addressing service

delivery bottlenecks that limit quality in essential neonatal care in low income countries like Kenya.(27)

From Table 3.2, the NTISS appeared to be the most promising in this regard as it predicts both in-hospital mortality (useful for clinical decisions) and resource use (useful for service planning).(104) In addition, previous work in Kenyan neonatal units has demonstrated the availability of data on treatments (therapeutic parameters) from routine data sources.(32) Therefore, I proceeded to conduct a systematic review of all models using therapeutic parameters in neonatal medicine. The specific review objectives are outlined in Section 3.1.3.

### *3.1.3 Objectives of the systematic review*

The aim of the second step towards identification of a candidate model suitable for low income countries was therefore to systematically identify and characterise prediction research that has used *therapeutic parameters* (hereby referred to as treatment intensity scores with reference to the NTISS) in neonatal care and specifically to:

1. Assess the certainty of evidence on predictive performance of existing neonatal treatment intensity scores in predicting clinical outcomes and resource utilisation.
2. Assess the applicability of existing neonatal treatment intensity scores to essential neonatal care in low income countries.

## 3.2 Methods

### *3.2.1 Overview of methods*

The review was done in accordance with the recently published **CH**ecklist for critical **A**ppraisal and data extraction for systematic **R**eviews of prediction **M**odelling **S**tudies(CHARMS). This provides guidance for systematic reviews of prediction model studies.(101) In addition, The **G**rading of **R**ecommendations **A**ssessment, **D**evelopment, and **E**valuation (GRADE) was used as a guiding framework to assess the quality of the retrieved articles.(108) A detailed description of the CHARMS checklist and how GRADE was utilised is provided in Sections 2.2.2 and 2.2.3 respectively of Chapter 2. Reporting of the review was done using the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) recommendations.(190)

Details of the protocol for this systematic review were registered on the International Prospective Register of Systematic Reviews (PROSPERO) and can be accessed at:

http://www.crd.york.ac.uk/PROSPERO/display_record.asp?ID=CRD42016034205.

### *3.2.2 Eligibility criteria*

Inclusion and exclusion criteria are presented in Table 3.3.

***Table 3.3**: Eligibility criteria for inclusion in the systematic review*

| Criteria | Inclusion | Exclusion |
|---|---|---|
| **Model type** | Prognostic models | Diagnostic models* |
| **Intended scope of the review** | Inform decision making at individual (estimated risk of outcome) level and for planning service delivery | |
| **Types of modelling studies** | Development and/or validation | |
| **Target population** | Neonates admitted to a neonatal unit in any country | Older children or adults |
| **Predictors** | Any use of treatments or interventions as predictors | |
| **Outcomes** | Any outcome | |
| **Time of prediction** | No restriction | |
| **Intended moment of use** | No restriction | |

* Diagnostic models estimate probability that a specific disease is currently present in an individual in contrast to prognostic models that estimate probability of future events.(92)

### *3.2.3 Information sources*

The primary electronic data base used was PubMed with supplementary searches conducted in EMBASE (OVID), CINAHL, Global Health Library (Global index, WHO) and Google Scholar. The last search conducted was on 21st December 2016. Bibliographies of identified papers were also hand searched for additional papers.

*3.2.4 Search strategy*

The following key search terms were used “neonate” or “neonatal or newborn”, “treatment or therapeutic” or “therapy”, “intensity” and “score or scoring”. In PubMed the search strategy was implemented using medical subject headlines (MeSH) where applicable and the appropriate Boolean terms;(((((("Neonatology"[Mesh]) OR "Infant, Newborn"[Mesh])) AND (("Therapeutics"[Mesh]) OR "therapy"[Subheading])) AND intensity) AND scor*.

The primary search was augmented by substituting a validated search string for finding prediction models in PubMed for the scor* term used in the first search string ; ((Validat* OR Predict* OR Rule*) OR (Predict* AND (Outcome* OR Risk* OR Model*)) OR ((History OR Variable* OR Criteria OR Scor* OR Characteristic* OR Finding* OR Factor*) AND (Predict* OR Model* OR Decision* OR Identif* OR Prognos*)) OR (Decision* AND (Model* OR Clinical* OR Logistic Models/)) OR (Prognostic AND (History OR Variable* OR Criteria OR Scor* OR Characteristic* OR Finding* OR Factor* OR Model*))).(165)

A similar approach was used in all the other electronic databases the details of which are provided in Appendix 7. This search strategy was developed with input from an expert medical librarian.

### *3.2.5 Study Selection*

All articles identified by the search were initially screened for eligibility on title independently by two reviewers with disagreements resolved by discussion. The search results were exported to the reference management software EndNote X7 (Thomas Reuters, Philadelphia, USA). Duplicate articles were eliminated and the remaining titles screened for relevance. Abstracts of relevant articles were screened against predefined inclusion and exclusion criteria (Table 3.3) and a short list created. The full texts of the resulting articles were retrieved and similarly assessed for eligibility using the predefined criteria.

### *3.2.6 Data Collection Process*

Two independent reviewers extracted data using the CHARMS checklist and any disagreements were resolved by discussion.

### *3.2.7 Data Items*

Two independent reviewers extracted data using a standardised form based on the CHARMS checklist (Appendix 1) and any disagreements were resolved by discussion. Data elements extracted were: study design, participants, geographical location, outcomes predicted, description of model development (type of model, e.g. logistic regression), number and type of predictors, number of study participants and number of outcome events, handling of missing data and model performance (calibration, discrimination)

### *3.2.8 Critical appraisal*

Limitations in the conduct of individual studies were assessed using criteria defined by the CHARMS guidance (see Section 2.2.2 , Chapter 2) and tabulated.(101) In the absence of a specific tool for assessing risk of bias across studies for clinical prediction models the GRADE approach for diagnostic studies was used to guide the rating of certainty in the estimates of predictive performance reported.(111) Studies were initially rated as high quality if they were prospective cohort studies as recommended by CHARMS.(101) Subsequently, they were downgraded if there were serious limitations in the conduct of the study as defined by CHARMS, and if there was inconsistency, indirectness and imprecision as defined by GRADE (Section 2.2.3, Chapter 2). Certainty of the evidence on predictive performance was thus rated as high, moderate, low or very low for each outcome. Applicability is the extent to which the reviewed models are suitable for use in the target population. The intended population in this case is neonates in facilities providing essential neonatal care in LMICs. This was assessed using the CHARMS applicability criteria: study design, participant characteristics, outcomes predicted, predictors used and model performance.(101)

### *3.2.9 Synthesis of results*

No summary measures were computed as studies were judged unsuitable for metanalysis due to heterogeneity in the conduct and reporting. Results were therefore summarised descriptively and synthesized using a narrative approach.

### 3.3 Results

#### *3.3.1 Study Selection*

A total of 3249 unique articles were identified by the search strategy, of which 3229 were excluded based on the title and abstract (Figure 3.1).

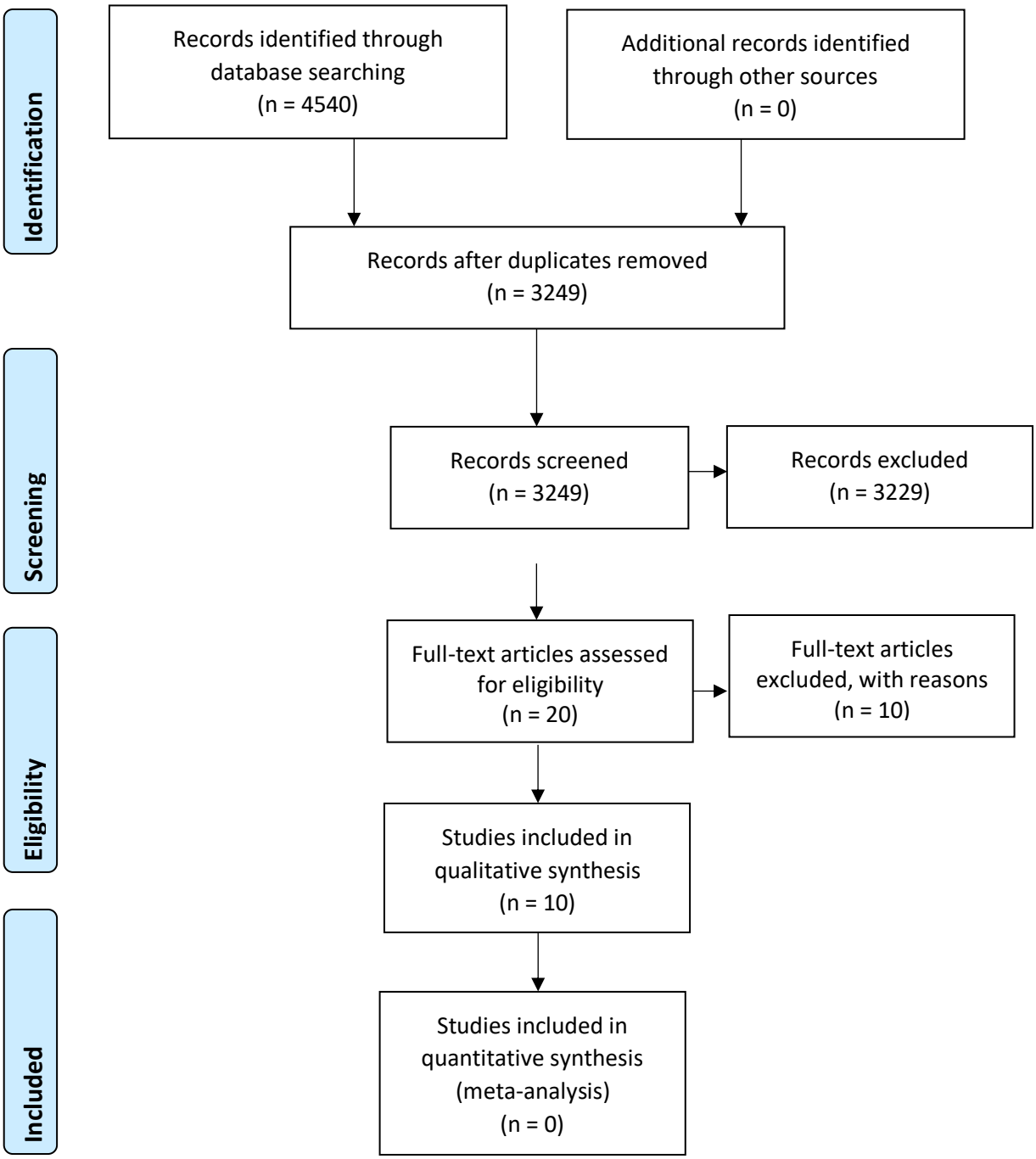



***Figure 3.1****: Flow diagram for the selection of studies on neonatal treatment intensity models*

The full texts of twenty articles were screened, of which 10 articles met the inclusion criteria and were included in this review (Figure 3.1). Articles were excluded for the following reasons; a population other than neonates was studied, intensity referred to a non-therapeutic intervention (e.g. radiological imaging intensity), treatment intensity was measured by means other than a score (e.g. proportion of days that hospital intensive care was required) and studies limited to neonates with congenital anomalies considered to be lethal.

### *3.3.2 Study characteristics*

Table 3.4 summarises the general characteristics of the included studies. All the studies were conducted in tertiary neonatal intensive care units with 5 single centre and 5 multicentre studies including a total of 12,899 neonates. Sample sizes ranged from 22 to 9,978 with a median of 163. The largest study of 9,978 neonates included 23 neonatal units.(191) Two out of the ten articles reported score development, while eight were reports of external application of existing scores to new populations. Gray and colleagues developed the Neonatal Therapeutic Intervention Scoring System; NTISS while Shah et al developed a score using selected variables derived from the NTISS.(104, 191) Seven of the eight articles reporting external application to new populations used the NTISS.(192-198) The eighth used a therapeutic score originally developed for adults the Therapeutic Intervention Scoring System; TISS.(199) A summary of data extracted from each article is presented in Table 3.8.

***Table 3.4**: Characteristics of studies included in the review*

| Study | Study Dates | Design | Setting | Sample size | Mortality* |
|---|---|---|---|---|---|
| **Georgieff,1989** | 1987 | Prospective cohort | Single academic centre NICU, USA | 55 | 0 |
| **Gray,1992** | 1989-1990 | Prospective & Retrospective | 3 academic centre NICUs, USA | 1768 | 114 |
| **Davies,1995** | Not reported | Retrospective | Single academic centre NICU, South Africa | 50 | 8 |
| **Eriksson,2002** | 1991-1995 | Retrospective | 2 NICUs, Sweden | 240 | 39 |
| **Zupancic,2002** | July 1998 & August 1999 | Prospective | Single academic NICU, USA | 154 | Not reported |
| **Mendes,2006** | 2004 | Prospective | 2 NICUs, Brazil | 96 | 9 |
| **Rojas,2011** | April 2007 | Prospective | Academic centre, NICU and intermediate care unit, Colombia | 22 | Not reported |
| **Oygur,2012** | 2006-2010 | Prospective & Retrospective | Single, academic centre NICU, Turkey | 364 | 103 |
| **Shah,2015** | 2010-2012 | Retrospective | 23 tertiary NICUs, Canada | 9978 | 650 |
| **Wu,2015** | 2007-2011 | Retrospective | Single academic NICU, Taiwan | 172 | 18 |

*Whereas only mortality is presented here, data was extracted on all outcomes reported by the authors

† Neonatal intensive care unit

### *3.3.3 Critical appraisal of individual studies*

#### *a) All studies*

All the included studies were *individually* assessed for limitations using the CHARMS criteria (Table 2.3, Chapter 2) in two steps: limitations in the design and conduct of the study (Table 3.5) and limitations in statistical analyses (Table 3.6). With regards to study design and conduct, four were prospective cohort studies as shown in Table 3.3. From Table 3.5, most of the studies (8/10) recruited all eligible infants and therefore were deemed to have no selection bias. In addition, none of the articles explicitly reported blinding during collection of data on predictors and outcomes and were therefore judged to have risk of bias. Missing data may also

give rise to bias during analysis and only one study was judged to have no risk of bias due to missingness.(191)

***Table 3.5**: Limitations in individual studies with respect to study design and data collection*

| Study | Study type | Participants | Outcome(s) | Predictors | Sample size | Missing data |
|---|---|---|---|---|---|---|
| **Georgieff ,1989** | No limitation<br>• Prospective cohort | No limitation<br>• All infants recruited | Limitation<br>• Risk of measurement error<br>• No blinding | Limitation<br>• Risk of measurement error<br>• No blinding | Limitation<br>• Risk of overfitting | Limitation<br>• Missingness not reported |
| **Gray ,1992** | No limitation<br>• Prospective cohort | No limitation<br>• All infants recruited, Medicus* for one hospital only | Limitation<br>• Risk of measurement error<br>• No blinding | Limitation<br>• Risk of measurement error<br>• No blinding | Limitation<br>• Not possible to assess | Limitation<br>• Risk of bias due to missingness |
| **Davies ,1995** | Limitation<br>• Retrospective cohort | Limitation<br>• No information on sampling | Limitation<br>• Risk of measurement error<br>• No blinding | Limitation<br>• Risk of measurement error<br>• No blinding | Limitation<br>• Risk of overfitting | Limitation<br>• Missingness not reported |
| **Eriksson ,2002** | Limitation<br>• Retrospective cohort | No limitation<br>• All infants recruited | Limitation<br>• Risk of measurement error<br>• No blinding | Limitation<br>• Risk of measurement error<br>• No blinding | Limitation<br>• Risk of overfitting | Limitation<br>• Risk of bias due to missingness |
| **Zupancic ,2002** | No limitation<br>• Prospective cohort | Limitation<br>• Risk of selection bias | Limitation<br>• Risk of measurement error<br>• No blinding | Limitation<br>• Risk of measurement error<br>• No blinding | Limitation<br>• Risk of overfitting | Limitation<br>• Missingness not reported |
| **Mendes ,2006** | No Limitation<br>• Prospective cohort | No limitation<br>• All infants recruited | Limitation<br>• Risk of measurement error<br>• No blinding | Limitation<br>• Risk of measurement error<br>• No blinding | Limitation<br>• Risk of overfitting | Limitation<br>• Risk of bias due to missingness |
| **Rojas ,2011** | No Limitation<br>• Prospective cohort | | Limitation<br>• Risk of measurement error<br>• No blinding | Limitation<br>• Risk of measurement error<br>• No blinding | Limitation<br>• Risk of overfitting | Limitation<br>• Missingness not reported |
| **Oygur,2012** | Limitation<br>• Retrospective cohort plus prospective cohort | No limitation<br>• All infants recruited | Limitation<br>• Risk of measurement error<br>• No blinding | Limitation<br>• Risk of measurement error<br>• No blinding | Limitation<br>• Risk of overfitting | Limitation<br>• Risk of bias due to missingness |
| **Shah,2015** | Limitation<br>• Retrospective cohort | No limitation<br>• All infants recruited | Limitation<br>• Risk of measurement error<br>• No blinding | Limitation<br>• Risk of measurement error<br>• No blinding | No limitation<br>• Not possible to assess | No limitation<br>• 1 patient excluded due to missingness |
| **Wu,2015** | Limitation<br>• Retrospective cohort | No limitation<br>• All infants recruited | Limitation<br>• Risk of measurement error<br>• No blinding | Limitation<br>• Risk of measurement error<br>• No blinding | Limitation<br>• Risk of overfitting | Limitation<br>• Missingness not reported |

Risk of bias assessed under each element of data extracted as guided by CHARMS
Presence of risk of bias is reported as a limitation
*Medicus: a proprietary nursing acuity tool used to estimate nursing workload (Medicus System, Evanston

*b) Score development studies*

In this section I now describe in detail the 2 studies that reported on the development of a treatment intensity score as seen in Table 3.6.(104, 191) The Neonatal Therapeutic Intervention Scoring System (NTISS) was developed to serve as a "therapy based severity-of-illness tool for use in intensive care "by modification of the adult Therapeutic Intervention Scoring System" (TISS).(104, 200) Shah and colleagues on the other hand developed a score to measure intensity of NICU resource use using information on 24 resource use items (many that are included in the NTISS).(191) In both of these studies treatment predictor selection and their respective predictor weights were determined by consensus amongst experts and in this aspect are similar to "medical" scores (see Section 3.1.2 (b)).

With respect to prediction of mortality, Table 3.4 shows there were a total number of 114 deaths (events) in Gray et al and 650 deaths in Shah et al.(104, 191) From a predictive modelling perspective it is recommended that there should be at least 10 outcome events (deaths in this case) for each predictor variable in the regression model.(101) It was however not possible to assess for risk of overfitting based on the sample size (Table 3.5) since none of the two studies published the final regression models and therefore it is not clear what the number of events per predictor (EPV) were.

In statistical analysis, none of these two score development studies specified a regression model that included the therapeutic interventions as individual predictors. Rather, the predictor weights were assigned sub-scores by an expert panel which

were then summated to give a total score for each patient. The statistical relationship between the total scores and the outcomes was then examined (Table 3.5). Nonetheless, Shah et al reported model discrimination with a c-statistic of a regression model. The study describing the development of NTISS did not report model discrimination but assessed calibration using the Hosmer-Lemeshow (H-L) technique(Table 3.6).(104) Finally, neither of these two studies presented a final model or score chart to demonstrate how to determine predicted probabilities of the outcomes for individual patients.

***Table 3.6**: Limitations in individual studies with respect to statistical analysis*

| Study | Study type | Score performance† | Score validation‡ | Score presentation§ |
|---|---|---|---|---|
| **Georgieff,1989[199]** | Evaluation of existing score* | Limitation<br>• Not reported | Limitation<br>• Not reported | Limitation<br>• Not reported |
| **Gray ,1992[104]** | Development of new score | Limitation<br>• Calibration only (H-L) ‖ | Limitation<br>• Not reported | Limitation<br>• Not reported |
| **Davies ,1995[192]** | Evaluation of existing score* | Limitation<br>• Not reported | Limitation<br>• Not reported | Limitation<br>• Not reported |
| **Eriksson ,2002[193]** | Evaluation of existing score* | Limitation<br>• Discrimination only(AUROC) | Limitation<br>• Not reported | Limitation<br>• Not reported |
| **Zupancic,2002[194]** | Evaluation of existing score* | Limitation<br>• Not reported | Limitation<br>• Not reported | Limitation<br>• Not reported |
| **Mendes ,2006[195]** | Evaluation of existing score* | Limitation<br>• Not reported | Limitation<br>• Not reported | Limitation<br>• Not reported |
| **Rojas ,2011[196]** | Evaluation of existing score* | Limitation<br>• Not reported | Limitation<br>• Not reported | Limitation<br>• Not reported |
| **Oygur,2012[197]** | Evaluation of existing score* | Limitation<br>• Discrimination only (AUROC) | Limitation<br>• Not reported | Limitation<br>• Not reported |
| **Shah et al,2015[191]** | Development of new score | Limitation<br>• Discrimination only (c-statistic) | Limitation<br>• Not reported | Limitation<br>• Not reported |
| **Wu ,2015[198]** | Evaluation of existing score* | Limitation<br>• Discrimination only (AUROC) | Limitation<br>• Not reported | Limitation<br>• Not reported |

Study limitations were assessed under each element of data extracted as guided by CHARMS
* None of these applied a regression formula from the original score development study to the new population but instead specified new models thus were in effect model re-development rather than external validation studies
† Performance: score discrimination or calibration
‡ Validation: Internal validation (to quantify model overfitting) or external (performance in new population)
§ Presentation: regression formula or score chart
‖ H-L: Hosmer-Lemeshow test

*c) Application of existing scores to external populations*

Here, I now present the studies that applied existing scores to new populations. Eight articles described the evaluation of existing therapeutic intensity scores in a different geographical setting but still in neonatal intensive care (Table 3.5). Seven of these evaluated the NTISS.(192-198) The eighth evaluated a version of the adult TISS in neonates.(199, 201) For external validation studies using regression approaches, sample sizes between 100 and 250 outcome events have been suggested.(202-204) With respect to the primary outcome (mortality) only 3/8 studies had at least 100 outcome events (Table 3.4).(191, 197, 205)

The aim of validation of prognostic models is to examine if they work as well in populations other than those they were derived in.(206) The key step is the application of the exact model (formula) obtained in the score development work to the external population to predict the probability of the outcome in this new data set and thus allow comparison to the observed outcome.(101, 207) This is a vital test as the model may perform poorly when applied to new populations.(207) Zupancic et al entered individual treatments as separate parameters in a linear regression model with personnel time as outcome.(194) Oygur et al also had individual treatments as separate parameters but didn't specify how they computed the predicted in-hospital mortality.(197) In contrast, the other six studies summated the treatments to give a single score which was then used in subsequent analyses as a single parameter.(192, 193, 195, 196, 198, 199) This was the same approach used by Gray et al in developing the NTISS.(104)

In terms of model performance, Eriksson et al, Oygur et al and Wu et al computed discrimination using area under the receiver operating characteristic curve (AUROC) but none of the studies reported model calibration.(193, 197, 198) From the foregoing, none of these 8 studies applied the exact formula as obtained from the original score development work but instead conducted fresh analyses, these were therefore all in effect model re-development studies in a new population rather than external validation.

*3.3.4 Critical appraisal of certainty of the evidence in score performance for outcomes across all studies*

In the preceding section we looked at the limitations in the *individual* studies identified. In this section I first give an overview of prediction model performance metrics then go on to describe the assessment of the certainty in the evidence of model performance in predicting specific outcomes *across all* retrieved studies.

*a) Overview of prediction model performance measures*

The essence of prediction models is their performance in *estimating the probability of occurrence* of the outcomes of interest. This is quantified by discrimination and calibration.

Discrimination is defined as the probability that the *predicted risk* of the outcome will be higher for an individual who experiences the outcome (outcome observed) compared with one who does not (outcome does not occur).(152) Two metrics may be used in this regard: the area under the receiver operator characteristic curve (AUROC) and the C-statistic. These will range from 0.5 (no discriminative ability) to 1.0 (perfect discrimination).(152) Values greater than 0.80 are considered adequate.(153)

Calibration quantifies the extent to which the predicted probabilities for an outcome of interest are in line with the actual probabilities.(152) This is preferably done using a calibration plot but is more commonly yet erroneously determined using the H-L test

which is a goodness-of-fit test for logistic regression models. The H-L test compares the observed and expected outcomes using a $\chi^2$ test which does not indicate the direction of any mis-calibration and the p-value computed is dependent on both mis-calibration and sample size in contrast to the calibration plot which provides both the direction and magnitude of any mis-calibration.(135) In addition, the test has inadequate power to detect overfitting of the effects of the predictors.(136)

Besides predictive performance another aspect of the performance of statistical models is explanatory power. Shmueli defines explanatory modelling as “the use of statistical models for testing causal explanations.(208) The focus of this approach is in testing hypotheses regarding the effect of different exposures on the outcome of interest by estimating statistical associations. Explanatory power is thus distinct from predictive performance. However, the two approaches are often conflated creating ambiguity.(208) For this reason we also report explanatory performance of the treatment intensity scores as presented by the authors in order to make this distinction clear.

*b) Appraisal of certainty in predictive performance across the studies*

The outcomes were reported in 3 main categories: mortality, morbidity and resource utilisation (nursing workload, length of stay, time inputs, comparison of resource use). GRADE criteria (Table 2.6, Chapter 2) were used to assess the certainty in the evidence of predictive performance for each outcome *across all* studies where it was reported and tabulated in Table 3.7 below.

***Table 3.7**: Certainty in the evidence of predictive performance across all studies*

| Outcome | Studies (n) | Factors that may decrease certainty of evidence | | | | | Overall Certainty | Importance |
|---|---|---|---|---|---|---|---|---|
| | | Limitation | Indirectness | Inconsistency | Imprecision | Reporting bias | | |
| **Mortality** | 5 (2594) | Serious | Serious | None | None | Unlikely | ⊕⊕OO low | **Critical** |
| **Morbidity** | 2 (295) | Serious | Serious | None | None | Unlikely | ⊕⊕OO low | **Critical** |
| **Morbidity & Mortality** | 2(10,218) | Serious | Serious | None | None | Unlikely | ⊕⊕OO low | **Important** |
| **Nursing workload** | 3 (317) | Serious | Serious | None | None | Unlikely | ⊕⊕OO low | **Critical** |
| **Hospital Costs** | 1 (1768) | Serious | Serious | None | None | Unlikely | ⊕⊕OO low | **Important** |
| **Length of stay** | 1(1768) | Serious | Serious | None | None | Unlikely | ⊕⊕OO low | **Important** |
| **Time inputs** | 1(154) | Serious | Serious | None | None | Unlikely | ⊕⊕OO low | **Critical** |
| **Compare resource use** | 1(96) | Serious | Serious | None | None | Unlikely | ⊕⊕OO low | **Important** |
| **Rehabilitation** | 1 study (n=240) | Serious | Serious | None | None | Unlikely | ⊕⊕OO low | **Not Important** |

There is low certainty in the evidence of predictive performance for these outcomes for low resource settings as all the studies had serious limitations (column three, Table 3.7) in their conduct (as determined by CHARMS criteria) in addition to indirectness (column four, Table 3.7) as they were all conducted in high-resource settings.

*3.3.5 Synthesis of results of estimated performance*

The previous section dealt with critical appraisal of the predictive performance of neonatal treatment intensity scores for mortality, morbidity and resource utilisation. I now present a synthesis of the actual performance estimates. Heterogeneity in study design, statistical analysis methods and performance metrics precluded metanalysis. We therefore did a narrative synthesis of model performance using the three outcome categories and performance measures reported as guiding themes.

a) Mortality

Five studies reported on in-hospital mortality (Table 3.7 and Table 3.8). Three of these computed discriminatory performance using area under the receiver operating characteristic curve (AUROC) analysis.(193, 197, 198) Calibration was reported in only one article.(104) Explanatory rather than predictive performance of the summated treatment score for in-hospital mortality was reported in three articles.(104, 192, 198)

In Eriksson et al, AUROC for in-hospital mortality was 0.82 ± 0.04.(193) Oygur et al reported performance with 63 treatment variables and with sensitive treatments only (based on significant association with mortality by chi square test).The AUROC for all treatment variables (63 treatments) by birth weight categories were : 500-1499g; 0.851 (95% CI 0.809 to 0.885) ,1000-1499g; 0.834(95%CI 0.781 to 0.878) and 500-999g; 0.749(95%CI 0.662 to 0.822). Using Student's t-test to compare AUROCs for all 63 variables versus 18 sensitive variables resulted in a significant change for the 500-999g category only; 0.749 vs 0.823(P=0.02).(197)

The NTISS was developed for use in the first 24 hours after admission and its performance after 24 hours was unknown. Wu et al reported serial scores in preterm infants weighing <1500g and reported AUROC at 24 hours; 0.913, 48 hours; 0.955 and 72 hours; 0. 958.Confidence intervals for the AUROC were not provided but using Tukey's honest significant difference (HSD) test, there was significant difference between the 24-hour average score when compared to both 48 and 72 hour average scores ($p<0.001$) There was no difference between the 48 and 72 hour scores ($p=0.391$). (198)

With regards to explanatory performance, Gray et al found that an increase in NTISS correlated with increasing risk of physician assessed risk of in-hospital mortality ($r=0.70$, $p<0.0001$) and actual in-hospital mortality (Chi square for trend = 158.3, $P < 0.001$).(104) Davies et al compared predicted mortality with actual mortality in three NTISS categories (labelled A, B and C) and found no difference A; 5% versus 0%, B; 20% versus 22%, C; 30% versus 37% (for predicted vs observed respectively).(192) However neither, the method of derivation of these three categories nor the method of statistical testing was reported.(192) In Wu et al average NTISS scores at 24 hours, 48 hours, and 72 hours after admission were all statistically significantly higher in the mortality group than in the survival group.(all $p <0.001$, two-sample t-test).(198)

***Table 3.8**: Summary of results from individual studies*

| Study | Predicting prognosis | Estimating resource utilization |
|---|---|---|
| **Georgieff et al,1989** | • Linear relationship between TISS and PSI; r=0.75, p<0.001<br>• Mean TISS decreased admission to discharge (<14-day stay;8.2±8.1 vs 3.1±2.9 and >14 day;16.9±8.1 vs 5.4±4.0, p<0.001)<br>• Higher mean TISS on day 1 and 5 in HMD vs TTN (16.2 ± 9.4/11.9±7.1 vs 9.2 ± 6.0/3.8 ± 1.5)<br>• Above median TISS (≥7) longer to adequate nutritional intake (5.1 ± 3.3 days' vs 1.9 ± 1.9 days, $p < .001$)<br>• Days of inadequate nutritional intake correlated with first-day TISS (r = .66, $p < .001$) | • Linear relationship TISS and NUMIS; r=0.72, p<0.001<br>• Mean TISS scores increased with increase in NUMIS classification (p<0.001)<br>• Mean TISS higher for level 3(NUMIS 7-10) care vs level 2 (NUMIS 5-6); 10.8±6.6 vs 3.2±1.4, p<0.001 |
| **Gray et al,1992** | • NTISS associated with physician assessment of mortality (r=0.70, p<0.0001)<br>• Increasing NTISS categories associated with increased in-hospital mortality ($\chi 2$ for trend = 153.3, p<0.001)<br>• All NTISS sub-scores except transfusion therapy independently contributed to mortality (p <0.001)<br>• Close agreement between observed and predicted mortality (H-L test) | • NTISS (in survivors) correlated with<br>○ length of stay (all hospitals); r=0.37, p<0.0001<br>○ Hospital charges (Hospital A); r=0.64, p<0.0001<br>○ Nursing workload (Medicus, Hospital A); r=0.69, p<0.0001 |
| **Davies et al,1995** | • In-hospital mortality concurred with the predicted range for each risk group (A; predicted 5% actual 0, B; 20% vs 20%, C;30% vs 37%) | • Not reported |
| **Eriksson et al,2002** | • NTISS prediction of; in-hospital mortality; AUROC 0.82±0.04, early adverse outcome; AUROC 0.78±0.03, morbidity at 4 years; AUROC 0.59±0.05, overall adverse outcome; AUROC 0.70±0.04 | • NTISS contact with rehabilitation; ROC Az 0.59±0.06 |
| **Zupancic et al,2002** | • Not reported | • Nursing time inputs; predicted 42% of the variability ($R^2 = 0.42$, p<0.0001)<br>• Nursing & physicians combined; 23% of the variability ($R^2 = 0.23$, p<0.0001)<br>• Respiratory therapists; 36% of the variability ($R^2 = 0.36$, p<0.0001) |
| **Mendes et al ,2006** | • Not reported | • Mean daily value of overall NTISS; gradual and progressive downward trend up to the 31st day, difference in curves after day 31 (increase in the public NICU; p<0.001) |
| **Rojas et al 2011** | • Not reported | • Total NTISS 5200 (SD: 70.10), mean 12.3 per measurement corresponding to 123 minutes of nursing interventions over a 6-hour period |
| **Oygur ,2012** | • AUROC for predicting mortality with all variables; 0.851 (95% confidence interval [CI], 0.809–0.885); group 1,0.834 (95%CI:0.781–0.878); group 2 ,0.749 (95%CI: 0.662–0.822); group 3<br>• AUROC NTISS only sensitive variables ;0.848 (95%CI: 0.806–0.883); group 1,0.821 (95%CI: 0.756–0.858); group 2 ,0.823 (95%CI: 0.744–0.886); group 3<br>• Increase in outcome significant only for group 3 patients for AUC NTISS all variables vs NTISS for only the sensitive variables (P = 0.02) | • Not reported |
| **Shah et al,2015** | • Crude and adjusted resource use on the day of admission associated with composite outcome<br>• Odds of the composite outcome higher for infants admitted to an NICU on day resource use in the highest quartile (1.28 95 C. I1.07-1.54), compared to day resource use in the lowest quartile (ref category 1) | • Not reported |
| **Wu et al,2015** | • Average NTISS scores at 24 hours, 48 hours, and 72hours higher in the mortality group (all p < 0.001)<br>• Overall NTISS scores at 24 hours, 48 hours, and 72 hours; 20.31± 5.6, 23.82± 6.2, 24.68± 5.8, (p < 0.001 one-way ANOVA).<br>• Logistic regression NTISS score at 48 hours' difference between mortality and survivors (p =0.047)<br>• AUC for NTISS at 24 ,48 and 72 hours were 0.913, 0.955 and 0.958, respectively in predicting mortality<br>• Optimal NTISS score at 48 hours for predicting mortality was ≥28.0, sensitivity 100%, specificity 81.2% (p < 0.001)<br>• Logistic regression model with gestational age and NTISS score at 48 hours AUC 0.989, sensitivity of 100.0% and specificity of 96.1%. | • Not reported |

*b) Morbidity*

There were various morbidities reported in 2 studies (Table 3.6). Predictive performance for morbidity was reported in the article by Eriksson et al only; here NTISS poorly predicted morbidity at 4 years of age with AUROC of 0.59±0.05.(193) but calibration was not reported.

Georgieff et al on the other hand reported explanatory performance only. They found a linear relationship between severity of physiologic instability (measured by Physiology Stability index; PSI) and the Therapeutic Intervention Scoring System (TISS) ($r = 0.75$, $P < 0.001$). In addition, the mean TISS at admission was significantly higher in infants discharged after 14 days (prolonged hospitalization) compared to infants hospitalized <14 days: mean (sd); 16.9 ± 8.1 vs 8.2 ± 8.1, t-test p value <0.001. Further, mean TISS scores on hospital days I and 5 were higher for severe lung disease (hyaline membrane disease, HMD) compared to mild disease (transient tachypnoea of the newborn, TTN): mean±sd; 16.2±9.4/11.9±7.1 vs 9.2±6.0/3.8±1.5 respectively, t-test p<0.001(Table 3.7). Finally, with regards to nutritional morbidity, infants with high day one TISS (TISS above the median of 7) required a significantly longer time to achieve adequate nutritional intake than did infants with low TISS (TISS below the median): mean ± sd; 5.1±3.3 days vs 1.9±1.9 days, t-test p< 0.001.(199)

*c) Composite of mortality and morbidity*

Predictive performance for the composite outcome of morbidity and mortality was reported in two articles. These were discriminative performance measured by; AUROC in Eriksson et al and c-statistic in Shah et al.(191, 193)

Eriksson et al found that for early adverse outcome (all-cause mortality or oxygen dependence, intraventricular haemorrhage (IVH)/periventricular leukomalacia(PVL) or retinopathy of prematurity (ROP)) the AUROC curve for NTISS was 0.78±0.03.(193) In Shah et al, the odds of the composite outcome (all-cause mortality,≥ grade 3 IVH, PVL, stage ≥3 ROP, oxygen dependence or stage ≥ necrotizing enterocolitis (NEC)) were higher for infants admitted to an NICU that on the day the infant was admitted had resource use (as measured by their new score derived from NTISS) in the highest quartile, compared with infants admitted to NICUs with resource use in the lowest quartile (c-statistic = 0.86 on logistic regression model that also included NICU size, NICU occupancy rate on day of admission, gestational age, small for gestational age, multiple births, outborn, caesarean delivery, SNAP II score >20 and mechanical ventilation on day of admission).(191)

*d) Resource utilisation*

All the four articles reporting on resource utilisation as an outcome estimated the extent to which resource use was explained by treatment intensity. This was done by computing the statistical association between resource use (measured in various ways) and treatment intensity. None reported predictive performance.

Nursing workload

Georgieff et al used the Nursing Utilization Management Information System (NUMIS) to measure nursing workload.(199) This is a score calculated daily to determine the amount of nursing care required by a patient based on their condition.(209) Zupancic et al on the other hand measured nursing workload by observing personnel time inputs in a time and motion study(194) whilst Gray et al used Medicus (Medicus System, Evanston IL) described as “a patient-based proprietary nursing acuity tool…used to estimate nursing workload”.(104) Lastly, Rojas et al measured nursing time utilised on patient interventions per 6 hour shift though it is not reported how this was measured.(196)

Georgieff et al found a significant linear relationship between TISS and NUMIS scores ($r = 0.72$, $p < 0.001$). In addition, TISS scores increased significantly with each increase in NUMIS classification ($p< 0.001$). In Gray et al, among inborn survivors, significant correlations were found between admission-day NTISS scores and admission day workload as measured by Medicus ($r = 0.69$, $P < 0.0001$). Zupancic et al on the other hand reported that treatment intensity explained 42% of the variability in nursing time inputs ($R^2=0.42$, $p<0.0001$) while Rojas et al found that the total NTISS score of 5200 (measured 425 times for 22 neonates) gave an average of 12.3 NTISS points per measurement which

corresponded to 123 minutes of nursing interventions over a 6-hour period. However it is not clear how this relationship between NTISS score and nursing time was computed.(196)

Hospital Costs

Gray et al measured hospital costs in one of three study hospitals by summating charges (excluding physician charges) as recorded in the hospital's discharge abstract .(104) There was correlation between admission-day NTISS scores and total hospital charges among inborn survivors($r = 0.64$, $p < 0.0001$). Further analysis using a multivariable linear regression model that included NTISS scores, transport and birth weight explained 60% of the variation in charges (multiple correlation coefficient = 0.77).(104)

Length of stay (LOS)

Across all the three study hospitals, Gray et al found a positive correlation between NTISS scores and LOS for survivors ($r = 0.37$, $p<0.0001$). A multivariable linear model with NTISS scores, birth weight, transport and 5-minute Apgar score for one of the hospitals explained 48% of the variation in LOS (multiple correlation coefficients of 0.70).(104)

## 3.4 Discussion

### *3.4.1 Performance of neonatal treatment intensity scores*

Neonatal treatment intensity scores predict mortality and morbidity. Discriminatory performance as measured by the area under the receiver operating curve is poorest for long term morbidity (0.59) and highest for mortality in infants weighing 1000-1499g for NTISS measured at 72 hours post admission (0.958).(193, 198) Calibration was reported by Gray et al only who found a close agreement between observed and predicted mortality by the Hosmer-Lemeshow test.(104) Using a variety of tests of statistical association rather than predictive performance treatment intensity was found also to be associated with mortality, morbidity and resource utilization.(104, 194, 196, 198, 199).

Discriminatory performance of neonatal treatment intensity scoring for predicting in-hospital mortality measured by AUROC ranges from 0.749 to 0.958.(197, 198) The recommended threshold for good discrimination is 0.8 (Section 3.3.5). The reported performance thus suggests that the treatment scoring approach performs well in distinguishing between neonates who die and those who don't in neonatal intensive care. This discriminatory ability is comparable to that of more commonly used neonatal predictive models as shown in Table 3.9.

***Table 3.9:*** *Performance of neonatal severity of illness scores based on physiological parameters*

| Score | Year | Development Cohort | Predictive Performance |
|---|---|---|---|
| **SNAP II*** | 2001 | all weights and gestation | Predictive performance not reported |
| **SNAP- PE II†** | 2001 | all weights and gestation | • Discrimination; for total NICU mortality AUROC 0.91±0.091<br>• Calibration; Hosmer-Lemeshow *p* value 0.90 |
| **CRIB II‡** | 2003 | • gestation<31 weeks<br>• birthweight <1500 g | • Discrimination; for total NICU mortality AUROC 0.92±0.01<br>• Calibration; Hosmer-Lemeshow p value 0.829, Cox's assessment (slope p=0·642 constant p=0·141) |
| **TRIPS-II§** | 2013 | all weights and gestation | • Discrimination; for total NICU mortality AUROC 0.87±0.33, for seven day NICU mortality 0.90±0.30<br>• Calibration; Hosmer-Lemeshow, p value=0.14 |

* Score for Neonatal Acute Physiology version 2
† Score for Neonatal Acute Physiology –Perinatal Extension version 2
‡ Clinical Risk Index for Babies version 2
§ Transport Risk Index for Physiological Stability version 2

The performance metrics in Table 3.9 are from the original model derivation studies (i.e. model development). Discriminatory performance for mortality measured by AUROC ranges from 0.87±0.33 for the Transport Risk Index for Physiological Stability version 2; TRIPS-II (Lee et al., 2013) to 0.92±0.01 for the Clinical Risk Index for Babies version 2; CRIB II (Parry et al., 2003). The application of the treatment intensity scoring approach in adult and paediatric intensive care preceded use in the neonatal population. However, none of these adults or paediatric studies reported the discriminatory performance of the scores.(200, 201, 210-212)

With regards to calibration, Gray et al reported good calibration in predicting for in-hospital mortality by the NTISS using the Hosmer-Lemeshow (H-L) test. This means the predicted deaths closely matched the observed deaths (Section 3.3.5).

Similarly, as shown in Table 3.9 the H-L statistic was reported for the physiologic scores Score for Neonatal Acute Physiology–Perinatal Extension version II (SNAP-PE II), CRIB II and TRIPS II in predicting mortality with all demonstrating good calibration ($p>0.05$).

There is however low certainty in this evidence of performance with reference to essential neonatal care in low resource settings for two reasons. One, there were serious limitations identified in the design and conduct of these studies such that even for high income countries the certainty in score performance would be low as well. Secondly these results are not directly generalisable to essential neonatal care in low resource settings as the studies were all conducted in neonatal intensive care in high income settings given the differences in range of treatments, staffing and case-mix.

Whereas the focus of this work was to investigate predictive performance of these scores, the retrieved articles also reported on explanatory performance. Treatment intensity was found to be associated with mortality, morbidity and resource utilisation.(104, 194, 196, 198, 199) This is consistent with findings in adult and paediatric intensive care for mortality.(200, 211) Similarly, treatment intensity was also found to be associated with nursing workload and could be used to determine the workload a typical nurse is capable of per shift in adult intensive care.(200, 212) Finally, treatment intensity also correlates with costs in adult intensive care.(200) The reporting of such analyses underscores the importance of clearly distinguishing predictive from explanatory aspects of models.(208, 213)

*3.4.2 Generalisability to essential neonatal care in low resource settings*

The second objective of this review was concerned with generalisability of the identified treatment intensity scores to essential neonatal care in low resource settings. Generalisability is assessed by evaluating the performance of previously developed models in a new population i.e. external validation (Section 2.6.5, Chapter 2). There were no studies of the treatment intensity score in low resource settings identified. The eight studies which applied previously developed treatment intensity scores to new populations were potentially external validation studies.(192-199) However, none of these applied model coefficients derived from the derivation studies to the new populations to compute the predicted probabilities of the outcomes to allow comparison with the observed outcomes. None of the studies therefore qualified as an external validation.

The use of treatments as predictors presents three potential limitations. To begin with, the treatments prescribed are dependent on the availability of resources and may not necessarily be an accurate reflection of the patient's actual requirements. Secondly, there may be variation in clinical practice which will be reflected in differences in treatments and thus potentially in the scores. Thirdly, the treatment predictors used as predictors in the identified studies reflect the neonatal intensive care study sites where mechanical ventilation and exogenous surfactant amongst other expensive and invasive interventions are available. These cannot be applied directly to populations receiving essential neonatal care in LMICs, for example with respiratory support limited to oxygen via nasal cannula. However, the treatment intensity approach has potential in the Kenyan context since there is a relatively

high degree of standardization of essential neonatal care in the first referral level facilities which mitigates to some extent these limitations.(17)

### *3.4.3 Limitations of the review*

The systematic review was guided by the CHARMS recommendations which provides a clear guide on extracting data and identifying limitations in individual studies.(101) However, it does not provide guidance on how to assess the quality of evidence on individual outcomes across all the identified studies. As a result, the GRADE approach was used as a guiding framework for this purpose. While there is no GRADE guideline specifically for clinical prediction models, GRADE for diagnostic models was judged sufficiently similar and used for this review.(115, 124) A key strength of the GRADE approach is that the judgements on the quality of evidence are made on explicit criteria. Nonetheless, the PROBAST tool which is specific for this purpose is under development.(110)

### *3.4.4 Conclusion*

Existing neonatal treatment intensity scores show promise in predicting mortality and morbidity but there is low certainty in the evidence on their performance in essential neonatal care in low income countries (LICs). However, over 12,000 patients were included in this review of 10 studies in comparison to 467 in the only score (SAW) developed to date for LICs.(103) In addition, increase in treatment intensity is associated with higher nursing workload, hospital costs and length of stay. The approach may therefore be usefully developed further for low resource settings e.g. Kenya because treatment data may be easier to obtain compared to other parameters like measures of physiological status. Prognostic information

thus obtained can support decision making in the planning and organisation of quality essential neonatal care.

# Chapter 4 : Case fatality ratio, length of stay and time to in-hospital death or discharge

## 4.1 Introduction

### *4.1.1 Background*

In Chapters one to three, I have highlighted the key role that hospitals are expected to play in the reduction of neonatal mortality in LMICs. This means in-hospital mortality will play an increasingly bigger role as an indicator of progress in the improvement of neonatal survival in these countries. In light of this, in this Chapter I focus on two key outcomes, in-hospital mortality and length of stay following admission to the study neonatal unit. These cover the first objective of the thesis which aims to describe the clinical characteristics of the study population and in-hospital survival.

Health care outcomes in general are multidimensional and complex. Kozma and colleagues have proposed a framework of outcomes to illustrate this complexity; The Economic, Clinical, and Humanistic Outcomes (ECHO) model shown in Figure 4.1.(214, 215)

**The image originally presented here cannot be made freely available via ORA because of copyright. The image was sourced from** *Velentgas P, Dreyer NA, Nourjah P, Smith SR, Torchia MM. Developing a Protocol for Observational Comparative Effectiveness Research: A User's Guide. Agency for Healthcare Research and Quality; 2013*

***Figure 4.1:*** *The Economic, Clinical and Humanistic Outcomes model*

Reproduced from Velentgas P, Dreyer NA, Nourjah P, Smith SR, Torchia MM. Developing a Protocol for Observational Comparative Effectiveness Research: A User's Guide. Agency for Healthcare Research and Quality; 2013

In this model, clinical outcomes are defined as "Medical events that occur as a result of disease or treatment". Clinical outcomes data are a crucial component of efforts to develop and improve health systems globally.(55) Strengthening of these health systems is essential if the SDGs are to be achieved.(216) However, for clinical outcomes data to be effective in this role, the outcomes measured and reported must be those required by decision makers.(217) An example is the use of inpatient mortality as a performance indicator for neonatal care in LMICs. This is well recognised in the global neonatal health agenda with the Every Newborn Action Plan (ENAP) explicitly stating under strategic objective five (count every newborn): "measurement of outcomes is necessary to enable managers to improve performance".(218) These include data on deaths, causes of death, morbidity and disability.(218) Neonatal mortality is also included in the WHO 100 core health

indicators (2015) prioritised for reporting progress in the health sector nationally and globally.(219)

Prognosis may be defined as the probable course and outcome of a health condition over time.(220) Prognostic outcomes are therefore a type of clinical outcome whose importance is described under the PROGRESS theme one (Fundamental Prognosis) in Chapter 1.(68) Fundamental prognosis is therefore the theme for this Chapter illustrating the potential of this approach in understanding outcomes in neonatal units in LMICS. In-hospital mortality is the principal clinical outcome of interest. Understanding in-hospital mortality contributes towards improvement of clinical practice, strengthening health systems and patient (and family) experience.(221, 222) This will entail appreciation of the strengths and limitations of hospital mortality data as a performance indicator particularly in comparing facilities using standardised measures.(221)

In neonates, there is a well recognised pattern of time of death with most mortalities occurring in the first week of life and more so on in the first 24 hours.(1, 223, 224) In the hospital context, this makes length of stay an important related clinical outcome. The duration of hospital stay is in addition, an important determinant of resource use and is therefore a key consideration in service delivery planning and organisation in neonatal medicine.(225) In the next section, I describe the available data on in-hospital mortality and length of stay after admission to neonatal units in Kenya and other African countries.

*4.1.2 Case fatality and length of stay in neonatal units in Kenya and other African countries*

*a) Case fatality ratio*

In Chapter 1, we saw there is a paucity of data on neonatal in-hospital mortality in Kenya. In a country with an estimated birth cohort of ≈1.6 million (2016) and average health facility based delivery at 61% (226), I identified 15 published reports over a ten-year period (2007-16). Case fatality ratio ranged from 3% to 62 % (Table 1.8). This lack of data is compounded by the absence of neonatal data in the Kenya health information system which is hosted on the District Health Information System version 2 (DHIS2) platform.(227) In addition, at facility level in Kenya, there is currently no standard register for neonatal admissions. A similar scenario with regards to availability of data obtains in reports from other African countries as seen in Table 4.1.

***Table 4.1**: Case fatality ratio in neonatal units of African hospitals*

| Study | Country | Year | Hospitals | Sample size | Weight* (g) | Case fatality ratio |
|---|---|---|---|---|---|---|
| **Thomson et al** [228] | S. Sudan | 2017 | 1 | 4268 | All | 576/4268(14%) |
| **Ndombo et al** [229] | Cameroon | 2017 | 1 | 332 | All | 52/332 (16%) |
| **Abdallah et al**[230] | Uganda | 2016 | 1 | 1192 | All | 264/1192(22%) |
| **Kokeb et al** [231] | Ethiopia | 2016 | 1 | 325 | All | 75/325 (23%) |
| **Ndelema et al** [232] | Burundi | 2016 | 1 | 370 | All† | 66/370 (18%) |
| **Cavicchiolo et al** [233] | Mozambique | 2016 | 1 | 2118 | All | 544/2118(26%)‡ |
| | | | | 2158 | All | 396/2158(18%)‡ |
| **Tochie et al** [234] | Cameroon | 2016 | 1 | 703 | All | 124/703 (18%) |
| **Ballot et al** [235]§ | S. Africa | 2015 | 1 | 562 | 500-1501 | 149/562 (27%) |
| **Okomo et al**[236] | Gambia | 2015 | 1 | 4944 | All | 1734/494(35%) |
| **Hedstrom et al** [237] | Uganda | 2014 | 1 | 745 | All | 167/745 (22%) |
| **Ahlsen et al**[238] | Malawi | 2014 | 2 | 1496 | <2500 | 258/1496(17%) |
| **Chiabi et al**[239] | Cameroon | 2014 | 1 | 7824 | All | 769/7824 (9%) |
| **Ekwochi et al**[240] | Nigeria | 2014 | 1 | 261 | All | 37/261 (14%) |
| **Mah-Mungyeh** [241] | Cameroon | 2014 | 1 | 5828 | All | 583/5828(10%) |
| **Zash et al** [242] | Botswana | 2014 | 1 | 449 | All | 110/449 (25%) |
| **Zuniga et al** [243] | Burundi | 2013 | 1 | 453 | All | 67/453 (15%) |
| **Mmbaga et al**[244] | Tanzania | 2012 | 1 | 5033 | All | 536/5033(11%) |
| **Ersdal et al** [245] | Tanzania | 2012 | 1 | 256 | All | 47/256 (18%)‖ |
| **Pepler et al** [246] § ¶ | S. Africa | 2012 | 15 | 2307 | >401 | 91/2307 (3%) |
| **Shah et al**[247] | Eritrea | 2012 | 1 | 1502 | All | 87/1502 (8%) |
| **Onwuanaku et al** [248] | Nigeria | 2011 | 1 | 278 | All | 12/278 (4%) |
| **Ballot et al** [249] § | S. Africa | 2010 | 1 | 474 | <1501 | 140/474 (30%) |
| **Omoigberale et al**[250] | Nigeria | 2010 | 1 | 3075 | All | 625/3075(20%) |

* Birthweight
† Preterm only (<37 weeks)
‡ Pre- and post quality improvement intervention respectively
§ Vermont Oxford Network member neonatal units
‖ Follow up to 24 hours after delivery only
¶ All private facilities

Table 4.1 shows twenty-three reports from thirteen African countries published between 2010 and 2017 identified from PubMed (search strategy included in Appendix 8). Most of these studies (16/23) were conducted in tertiary hospitals with 21 out of 23 from a single facility. The largest number of centres included was 15.(246) Data were collected over different time periods ranging from three months to eleven years.(230, 244) This is reflected in the difference in sample size which ranges from 256 to 7824. There was variation in the case fatality ratio (CFR) with a median of 18% and range 3% to 35%. However, most of the studies (18/23)

reported CFR greater than or equal to 10%. The lowest CFR (3%) was observed in South Africa in the only study conducted in private sector facilities.(246)

*b) Time of in-hospital death and length of stay in the neonatal unit*

Time of in-hospital death following admission in the neonatal period varies from overall neonatal mortality at population level in three key ways: deaths may occur before admission or after discharge, case-mix varies between hospitals and in-hospital death includes deaths that occur after the neonatal period.

Out of the fifteen Kenyan articles described in Chapter 1 (Table 1.8), two reported time to in-hospital death and one reported on length of stay. Mwaniki et al found that 70% of the observed deaths occurred within 48 hours while Aluvaala and colleagues reported median time to death of 1 day (range <1-93).(25, 32) Regarding length of stay , Myhre et al reported median stay of 39 days (range 10–112) vs 28 days (range 16–54) before and after introduction of BCPAP for respiratory support.(251)

In the reports from other African countries, the time of in-hospital death was reported in 17/23 articles while 7/23 reported on length of stay. The proportion of deaths that occurred within 24 hours was described in 10/17 articles. This ranged from 18% to 67%.(239, 247) The proportion of deaths that occurred within the first week was reported in 9/17 studies and it ranged from 53% to 92%.(238, 244) Three studies reported median length of stay: 6 days (no range), 11days (range 1-171) and 5 days (interquartile range 2-15).(237, 242, 246) There was a variation in the mean length of stay reported in four studies: 26 days (24-28 days), 28 days (±22

days), 6 days (no standard deviation) and 5.5 days (±3.5 days).(231, 235, 247, 249)

These reports from neonatal units in Kenya and other African countries show a high case fatality ratio in general and especially for VLBW where specified. Where time to death was reported, most deaths occurred within the first week of admission with the highest number observed in the first 24 hours. Less than a third of these studies reported length of stay.

*4.1.3 Rationale*

From the foregoing, it is evident that there are few high quality data on in-hospital mortality and length of stay from neonatal units in Kenya and other African countries. Moreover, the lack of standardised reporting and likely unrepresentativeness makes comparison difficult. The available data shows in-hospital neonatal mortality is high and the number of deaths observed changes with time after admission. Given the crucial role hospitals are expected to play in neonatal survival, more and better data on mortality and length of stay are therefore required. Also required is better description of the populations seen in different hospitals to enable risk adjustment. A key example is criteria for admission to the neonatal unit as overall mortality may depend heavily on how many very low / low risk babies are admitted e.g. for observation.

Survival analysis approaches are well suited to examine both in-hospital mortality and length of stay to obtain insights beyond that provided by case fatality rates alone. However, as outlined in Chapter 2, the effect of competing risks must be

taken into account in order to obtain accurate estimates of the probability of being either discharged alive or an in-hospital death at various time points following admission. Hinchliffe and colleagues have demonstrated the utility of the competing risk approach for modelling length of stay where there are significant rates of mortality in the neonatal unit.(252)

Work by Gathara et al in Kenya and Duke et al in Papua New Guinea has shown that it is possible to use routine data in low resource settings to provide high quality estimates of in-hospital mortality in paediatrics.(253, 254) I therefore sought to use routine data from the largest (by patient volume) neonatal unit in Kenya to estimate in-hospital mortality and length of stay using competing risk survival analysis.

### *4.1.4 Objective*

The specific objective for this Chapter is to use routine data of patients admitted to a large urban neonatal unit in Nairobi, Kenya to:

i. Describe clinical characteristics of the patients
ii. Estimate case fatality ratio
iii. Describe length of stay
iv. Estimate the absolute probability of in-hospital mortality and discharge alive during stay in the unit

## 4.2 Methods

### *4.2.1 Study design*

This was a retrospective cohort study using a routine data set from a large urban neonatal unit. Study subjects were followed up from the time of admission to the unit to time of exit.

### *4.2.2 Setting*

The study site is the largest (by patient volume) neonatal unit in Kenya located in the largest public-sector maternity hospital in Nairobi, Kenya. A detailed description of this facility is provided in Chapter 2.

### *4.2.3 Participants*

The study participants were all the patients admitted to this large urban neonatal unit between April 2014 to December 2016. All the patients were entered into the data base consecutively at the point of discharge. In Chapter 2, I described the data set used in this Chapter spans admissions from April 2014 to December 2015. In this data set a 60% random sample of the observations had data on all the variables entered while all the observations had data on a restricted set of variables collected as illustrated in Figure 4.2

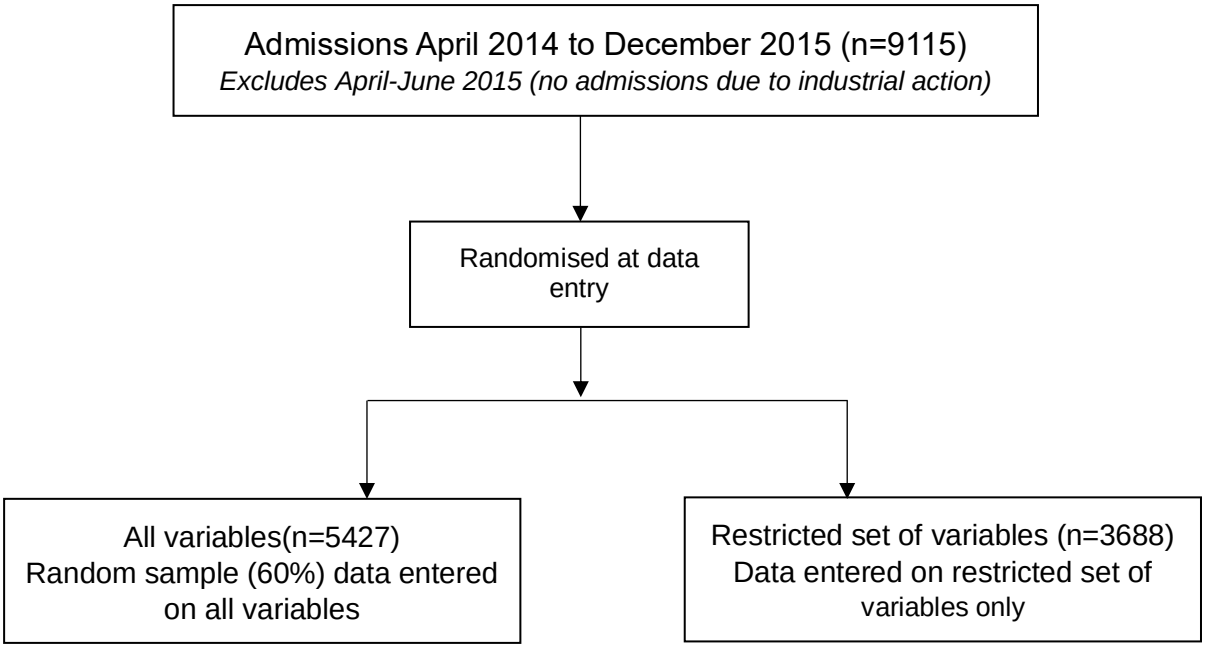



***Figure 4.2:*** *Selection of patients for inclusion in the study dataset*

### *4.2.4 Variables*

In accordance with the Chapter objective, the key variables are in two categories: clinical characteristics and survival analysis. The key clinical characteristics are: sex, birthweight, mode of delivery, place of delivery, admission diagnosis and outcome. The key variables in the survival analyses are outcome (described in section 4.2.9) and time to exit from the unit measured in days. All these variables are categorical except for time to exit from the unit which is a continuous variable.

*4.2.5 Data sources and measurement*

A detailed description of data collection procedures is provided in Chapter 2. Briefly, data were abstracted from patient records at the point of exit from the unit and entered directly into a REDCap$^{TM}$ data capture tool.(125)

*4.2.6 Bias*

Selection bias was minimised by including all patients admitted to the neonatal unit as described in Chapter 2 (Section 2.3). Patients were consecutively entered in the data base at the point of exit from the unit. Other data from this facility has shown that referral out of the neonatal unit is rare and would not occur directly from the maternity (yet to be published reports from the Nairobi Newborn Study).(82) This means that all sick neonates delivered in the hospital are admitted to the neonatal unit. In addition, data collection was conducted in the neonatal unit immediately after discharge (or death) to reduce the chances of missing patient records.

Information bias includes recall bias and observation bias. Recall bias is not a consideration in this case as no interviews were conducted. Observation bias was minimised by designing the electronic data capture tool to match the standard admission record (Neonatal Admission Record; NAR) and augmented by data quality assurance procedures described in Chapter 2 (Section 2.3). Nonetheless, there were missing data which could result in bias if associated in a systematic way with outcome. This issue of missingness is explored in detail (including multiple imputation) in Chapter 6.

*4.2.7 Study size*

A total of 9,115 subjects are included in the analyses in this Chapter. These were admissions to the neonatal unit from April 2014 to December 2015 but excluded a three-month period from April to June of 2015 when industrial action by healthcare workers in the public sector was underway. During this time, there were no admissions and therefore no data.

*4.2.8 Quantitative variables*

Birthweight is a quantitative variable but in this Chapter, it was handled as a categorical variable as it is in routine reporting of clinical characteristics of newborns. These are five categories: extremely low birth weight (ELBW, less than 1kg), very low birth weight (VLBW,1 to 1.499kg), low birth weight (LBW,1.5 to 2.499kg), normal (2.5-4kg) and macrosomia (greater than 4kg). Time to exit from the unit in days was handled as a continuous variable without any transformation.

*4.2.9 Statistical methods*

Clinical characteristics are presented using descriptive statistics only while survival analyses include both descriptive and inferential statistics. The clinical characteristics were summarised both numerically (proportions and 95% confidence intervals) and graphically (bar charts).

Two approaches were used in the survival analyses, Kaplan-Meier and Competing Risk survival analysis. Kaplan-Meier survival analysis was used to compute the probability of being alive in the neonatal unit over the duration of follow up. Follow up commenced at the point of admission and ended at the point of exit from the

unit. There were three possible outcomes at the time of exit: death, discharged alive and referred. In-hospital death was specified as the outcome of interest while discharge alive and referred were right censored. The probability of being alive and in the unit was computed at each time point where a death was observed. The risk set used to compute these probabilities at each of these points included those still alive in the unit and excluded those who had been discharged alive or referred. Kaplan-Meier survival curves were thus computed for the overall population, by sex and by birthweight. For the sex and birthweight curves, differences in the curves by category were tested using the Log-rank test. These analyses were implemented in R using the “survminer” package.(255)

A detailed comparison of the Kaplan-Meier and Competing Risks methods is presented in Chapter 2. The Kaplan-Meier method over estimates the probability of the outcome of interest in the presence of competing risks due to exclusion of the competing event from the risk sets. In this case, in-hospital mortality is the outcome of interest while discharge alive is the competing risk as it precludes the later observation of in-hospital mortality as outlined in Chapter 2. I used the cumulative incidence function to incorporate discharge alive in the risk set when computing the probability of in-hospital mortality over time. In these analyses, patients who were referred to other hospitals were right censored.(252)

Using the competing risk method, survival curves depicting absolute probability of in-hospital mortality and discharge alive were thus computed. This was done for the overall population, by sex and by birthweight. These analyses were

implemented in R using the “cuminc” function of the “cmprsk” package.(128, 134). The R code for all these analyses are included in Appendix 9.

## 4.3 Results

### *4.3.1 Patterns of admissions during the study period*

A total of 9,115 patients were included over the study period. Figure 4.3 illustrates the pattern of admission into the neonatal unit over the study period spanning from when data collection commenced in April 2014 to December 2015.There were 5463 admissions in 2014 and 3652 in 2015.

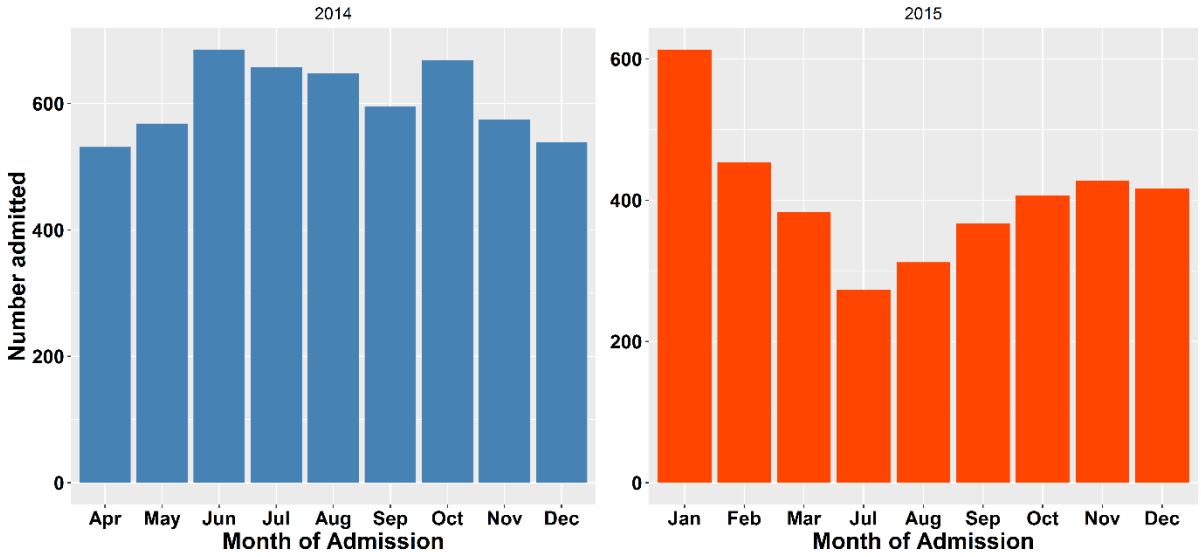



***Figure 4.3:*** *Pattern of admissions to the study neonatal unit*

These are data from 2014 (Left) and 2015 (Right). The 2015 data excludes April to June as there were no admissions to the unit following industrial action by health workers

There was month on month variation in number of admissions throughout the study period. However, a special cause of variation was observed in April to June 2015 during which industrial action by health workers was underway. This period is therefore excluded as there were no admissions to the unit. There was a gradual increase in number of admissions thereafter. The admissions by month peaked at ≈ 600 while the lowest number observed was ≈ 250.

*4.3.2 Characteristics of the study population and overall case fatality rate*

The clinical characteristics of the patients included are described in Table 4.2.

***Table 4.2**: Clinical characteristics of neonates included in the study data set*

| Characteristic | All patients (n=9115) | | Random sample (n=5427) | |
|---|---|---|---|---|
| | n | % | n | % |
| **1.Sex** | | | | |
| Male | 4966 | **55** | 2937 | **54** |
| Missing | 42 | **0.5** | 13 | **0.2** |
| **2.Birthweight (kg)** | | | | |
| <1 | 49 | **0.5** | 32 | **0.6** |
| 1-<1.5 | 280 | **3** | 136 | **2.5** |
| 1.5-<2.5 | 2132 | **23** | 1180 | **22** |
| 2.5-4 | 6287 | **69** | 3841 | **71** |
| >4 | 353 | **4** | 229 | **4** |
| Missing | 23 | **0.3** | 9 | **0.2** |
| **3.Mode of delivery** | | | | |
| Spontaneous vaginal | 5116 | **58** | 3011 | **57** |
| Assisted vaginal | 4 | **0.04** | 3 | **0.05** |
| Breech | 90 | **1** | 43 | **0.8** |
| Caesarean section | 3660 | **41** | 2212 | **42** |
| Missing | 245 | **3** | 158 | **3** |
| **4.Outborn** | | | | |
| Yes | 226 | **3** | 123 | **2** |
| Missing | 0 | | 0 | |
| **5. Apgar score (5 min)*** | | | | |
| 0-3 | | | 116 | **2.5** |
| 4-6 | | | 602 | **13** |
| 7-10 | | | 3992 | **85** |
| Missing | | | 17 | **0.4** |
| **7.HIV exposure** | | | | |
| Exposed | 547 | **6** | 319 | **6** |
| Missing | 529 | **6** | 305 | **6** |
| **8.Outcome** | | | | |
| Discharged alive | 8135 | **90** | 4839 | **89** |
| Dead | 839 | **9** | 507 | **9** |
| Referred | 108 | **1** | 62 | **1** |
| Missing | 33 | **0.4** | 19 | **0.4** |
| **9. Admission diagnosis†** | | | | |
| Birth asphyxia | 3867 | **42** | 2248 | **41** |
| Preterm/LBW | 2029 | **22** | 1107 | **20** |
| Neonatal sepsis | 960 | **11** | 550 | **10** |
| Respiratory distress syndrome | 724 | **8** | 371 | **7** |
| Neonatal jaundice | 513 | **6** | 312 | **6** |
| Others | 1014 | **11** | 649 | **12** |

* Data on Apgar score collected for random sample only

† These are admission episodes. Neonates with multiple diagnoses are counted under each of these making the total admission episodes greater than the study population

Table 4.2 demonstrates that the distribution of clinical characteristics in the full data and the sample data is comparable. There were slightly more males (4966/9115,

55%, 95% C.I 54 to 56%) than females in this population. Most of the neonates (6287/9115, 69%, 95% C.I 68 to 70%) were of normal birthweight (2.5-4kg) while the extremely low birthweight category (<1kg) had the least number of observations (49/9115, 0.5%, 95% C.I 0.4 to 0.7). With regards to mode of delivery, more than one in every three of babies admitted were born through caesarean section (3660/9115, 41%, 95% C.I 40 to 42%).

The key outcome of interest is in-hospital mortality and almost one in every ten babies admitted died in the unit (839/9115, 9.2%, 95% C.I 8.7 to 9.8%). Estimated gestational age at the birth was poorly documented with 70% missing and is thus not included in this table. Where documented, gestational age is determined from mother's recall of the last menstrual period or clinical estimation after birth as routine dating ultrasound is not available in this context. For the reported clinical characteristics, the highest proportion of observations with missing data was 6% for HIV exposure status.

*4.3.3 Length of stay in the neonatal unit*

*a) Overall length of stay*

Data on length of stay in number of days was available for 9107 out of all the 9115 neonates admitted. The distribution of length of stay was right skewed as demonstrated in Figure 4.4. This figure depicts the length of stay for all the neonates in the neonatal unit regardless of the outcome (death, discharge alive or referred) at the point of exit.

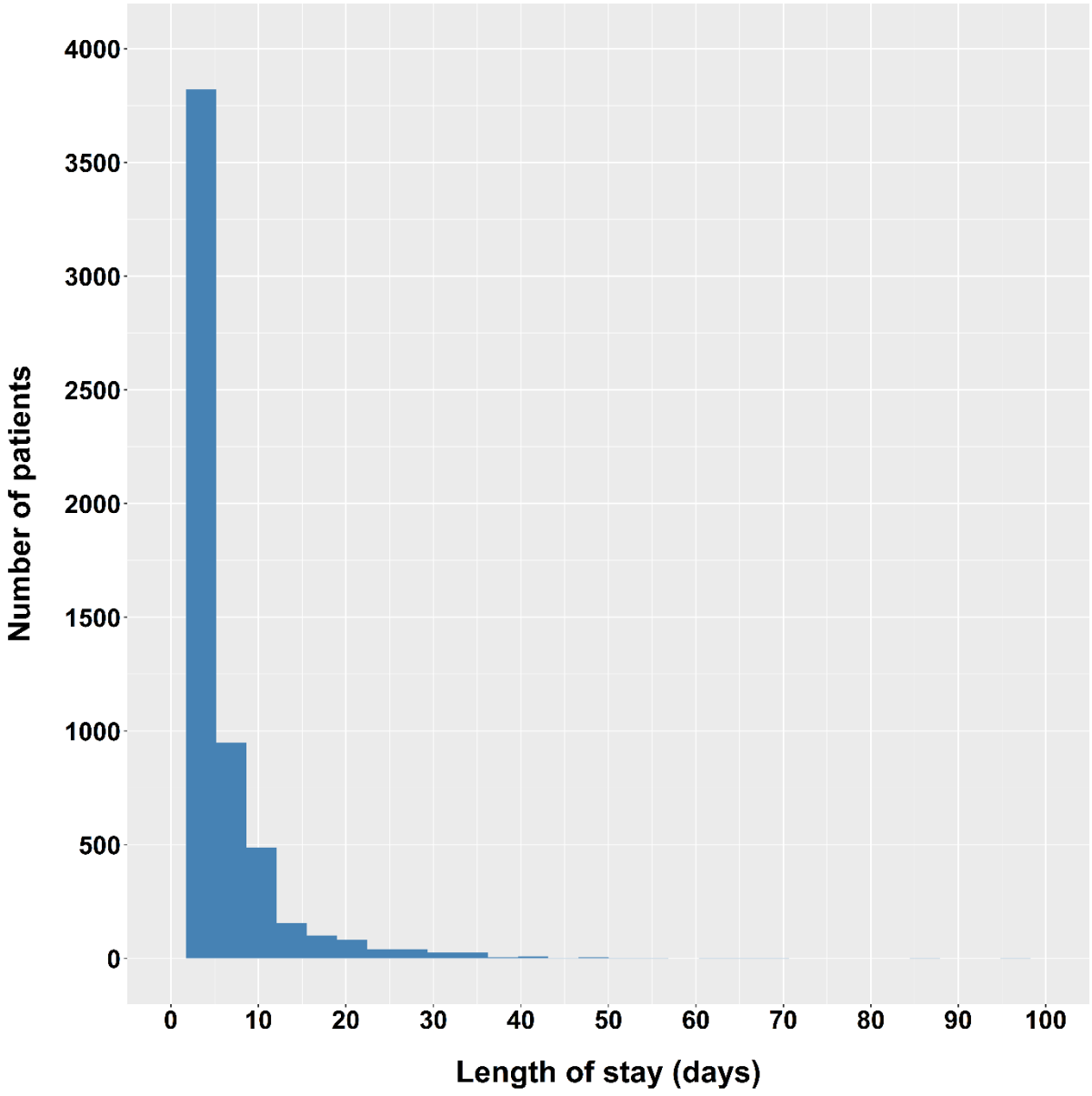



***Figure 4.4**: Distribution of length of stay in the neonatal unit for all the neonates*

From this histogram in Figure 4.4, we observe that length of stay is clustered around the first week (83%,7580/9107) following admission with a rapid drop in the number of observations thereafter. This is reflected in the median length of stay which was 2 days with interquartile range 1-5 days and range 0-98 days. In the next section, I describe the distribution of length of stay by birthweight categories.

*b) Length of stay by birthweight*

Data on length of stay and birthweight was available for 9092 out of all the 9115 patients admitted. These data are summarised in Figure 4.5. This boxplot describes the distribution of length of stay by five categories of birthweight for all patients regardless of outcome. The largest number of observations (6278/9115) had normal birthweight (2.5-4kg) while the least number of observations (49/9115) were in the extremely low birthweight (ELBW, <1kg) category.

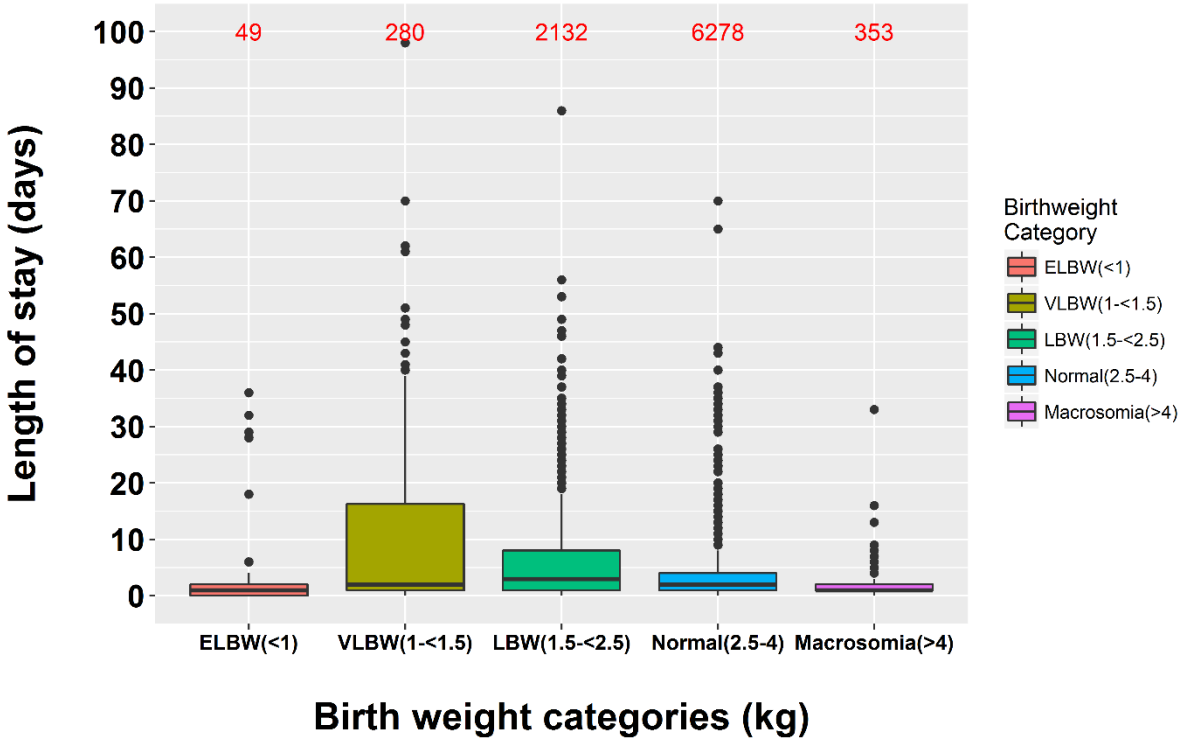



***Figure 4.5**: Length of stay in the neonatal unit by birthweight for all neonates*

Like Figure 4.4, Figure 4.5 shows the skewed distribution of length of stay across all the weight categories. The median stay is less than 5 days in all categories. The width of the interquartile range however varied, with the widest observed (1-17 days) for the VLBW. The shortest maximum length of stay was for Macrosomia (33 days) while the longest was for VLBW (98 days).

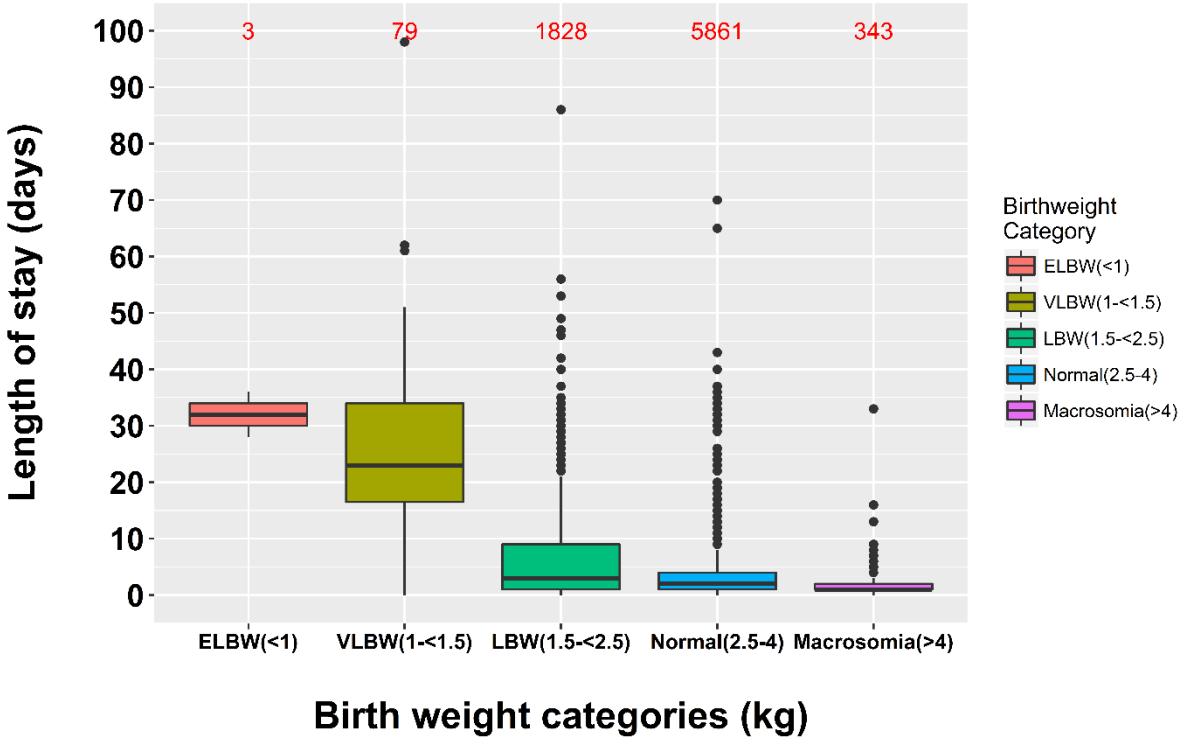



***Figure 4.6**: Length of stay by birthweight for neonates who survived to discharge*

Figure 4.6 shows length of stay for only those neonates who survived to discharge in contrast to Figure 4.5 which also includes those who died in-hospital. Figure 4.6 shows the small number of ELBW who survived to discharge (3/49). In addition, when compared to Figure 4.5, there is a marked upward shift in the median length of stay for the ELBW and VLBW; from 1 day to 32 days for ELBW and 2 days to 23 days for VLBW. This demonstrates the expected relatively long stays expected

for these two weight categories if they survive the early mortality observed in the first week following admission.

In Section 4.3.2 and 4.3.3 I have described the overall case fatality ratio and the length of stay. We have seen that almost 1 in 10 of neonates admitted to this neonatal unit during the study period died. In addition, length of stay was right skewed with most neonates spending less than a week in the neonatal unit but relatively long hospital stays of up to more than three months were observed in a few neonates. The relatively long length of stay is particularly pronounced in VLBW neonates who survive to discharge. In the next section, I will examine the time to in-hospital death using the Kaplan-Meier survival analysis method.

*4.3.4 Kaplan-Meier survival analysis of time to in-hospital death*

*a) Overall probability of being alive and in-hospital*

Data for 9074/9115 observations were included in the overall Kaplan-Meier survival analysis presented in Figure 4.7. This Figure shows the probability of a neonate being alive if they are still in the neonatal unit at different time points (i.e. probability of being event free over time). The graph describes survival over the total duration of follow up of 98 days.

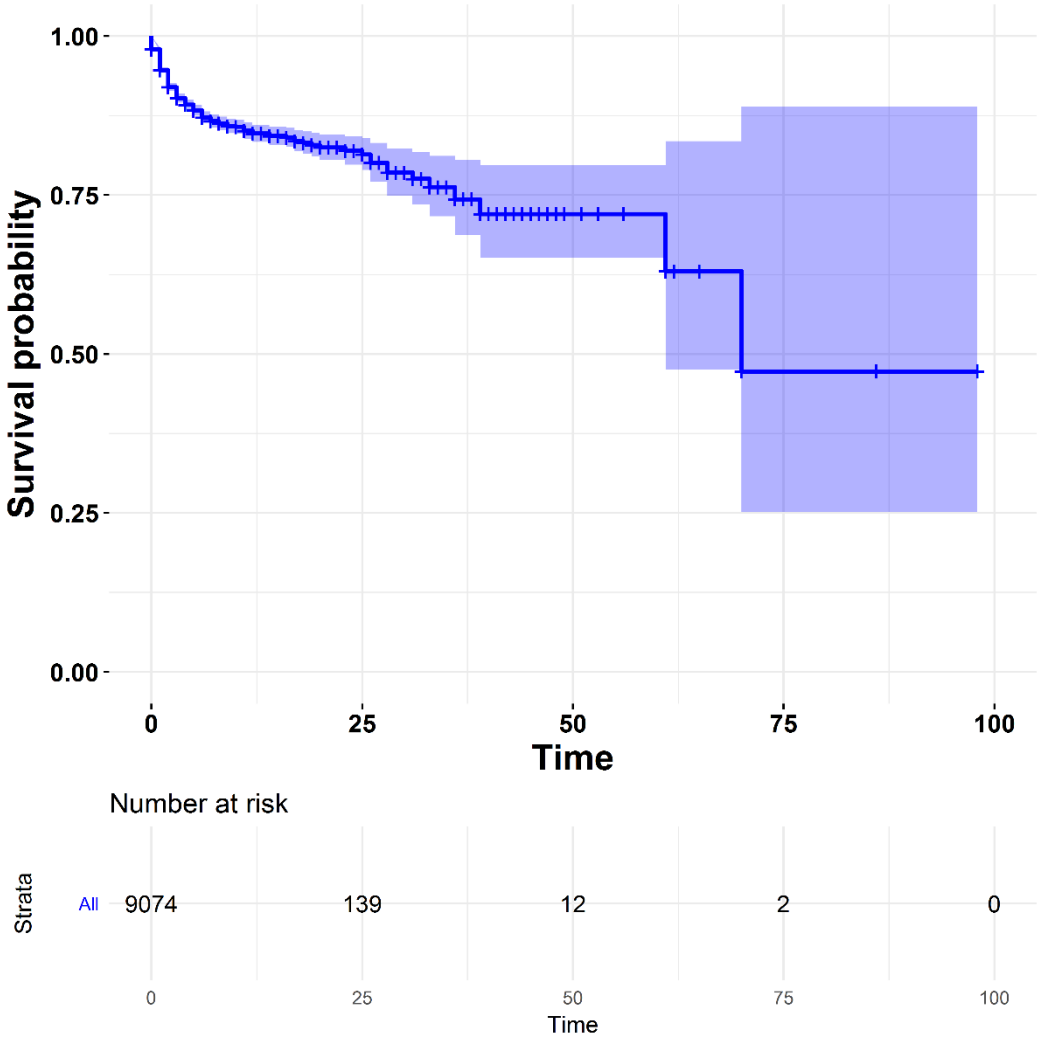



**Figure 4.7**: Probability of being alive while in hospital over time

In Figure 4.7, the occurrence of in-hospital deaths (in the neonatal unit) is marked by the steps in the survival curves. Those discharged alive and referrals were right censored and occurrence of these events is shown by the tick marks on the survival curves. There is a decline in survival probability from the first day (day 0) to day 70. On day 70, the probability of being alive if still in-hospital is 0.47. The converse of this is the probability of in-hospital death which is 0.53.

The reduction in the risk set over time explains the widening of the confidence interval of the survival curve with time as fewer individuals will still be in-hospital. This widening of the confidence interval due to reduction in sample size becomes particularly pronounced from roughly 25 days. This reflects the skewed distribution of length of stay described in Section 4.3.3. In the next two sections, I extend the Kaplan-Meier analysis to describe the pattern of in-hospital survival by sex and by birthweight.

*b) Probability of being alive and in-hospital by sex*

There were 4089 females and 4944 males with data on length of stay for a total of 9033 observations. Figure 4.8 describes the probability of a neonate being alive if they are still in the neonatal unit at different time points by sex. The maximum length of stay observed was higher in females at 98 days compared to 70 days for males.

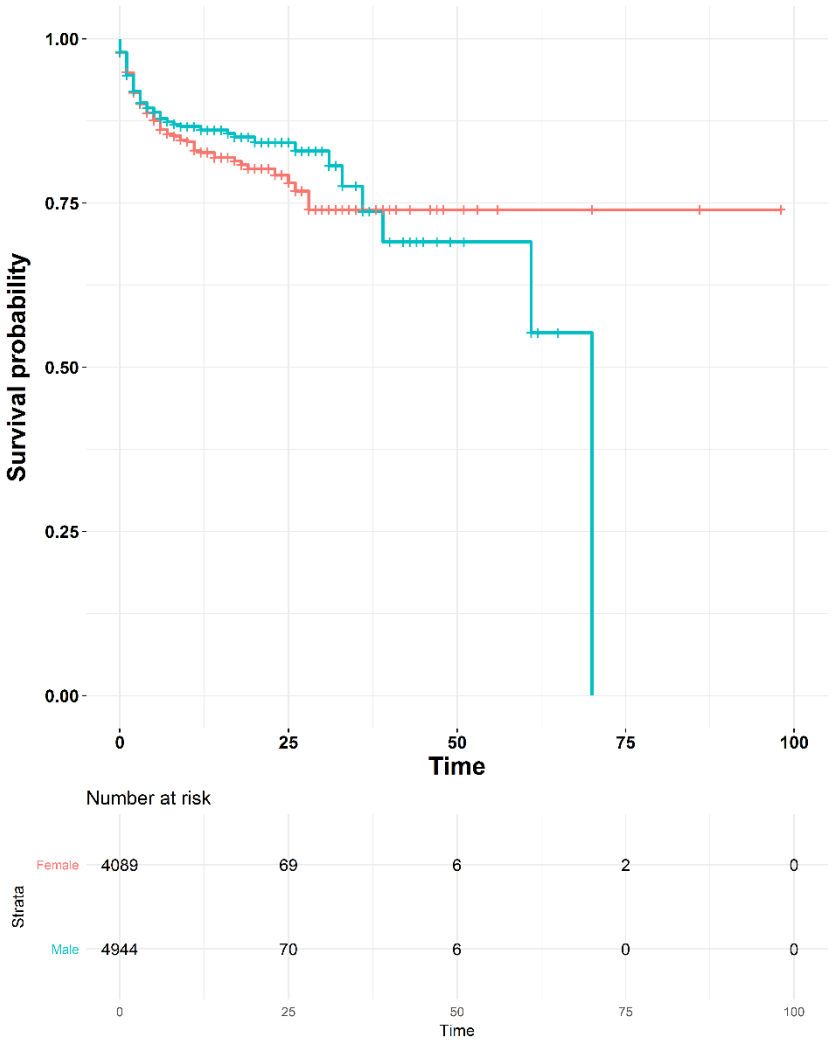



***Figure 4.8**: Probability of being alive while in hospital by sex*

Initially, the point estimates of probability of being alive in the neonatal unit are higher in males. Then, after five weeks the male survival curve is lower than the

female curve. The log rank test comparing the survival curve of male vs female yielded a chi square statistic of 1 with a p value of 0.135. This result means that there is weak evidence (p=0.135) against the null hypothesis which is no difference in survival between the two groups.

*c) Probability of being alive and in-hospital by birthweight*

There were 9052 observations available for examining in-hospital survival by birthweight. These data included, ELBW (49), VLBW (280), LBW (2118), Normal (6253) and Macrosomia (352). The pattern of survival is shown in Figure 4.9.

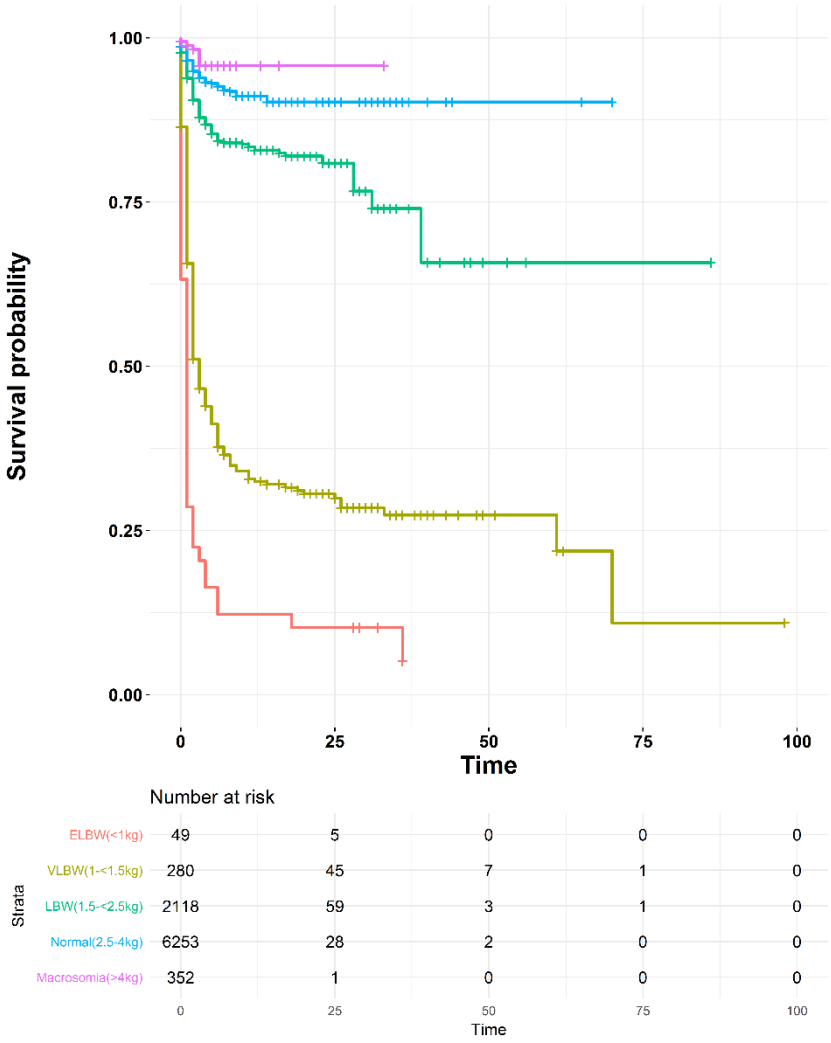



***Figure 4.9**: Probability of being alive while in hospital by birthweight*

As was seen in Figure 4.5, Figure 4.9 also demonstrates the variation in length of stay by birthweight category. The longest maximum length of stay was observed in VLBW (98 days) and the shortest in the Macrosomia category (33 days). The survival curves show that the probability of being alive while in the neonatal unit is highest for Macrosomia and lowest for ELBW. None of the five curves cross at any point suggesting that probability of being alive and in hospital over time is different for the birthweight categories. This is borne out by the log rank test; chi square statistic = 1417, $p<0.001$ which shows that the probability of being alive and in hospital over time differs by birthweight category.

In this section, I have described in-hospital survival following admission to the neonatal unit using the Kaplan-Meier approach. I have shown that the probability of being alive and in hospital declines over time. This differs by birthweight category but is not affected by sex. In the next section I will further examine in-hospital survival using the Competing Risk survival analysis method.

*4.3.5 Competing risk survival analysis*

Kaplan-Meier survival analysis as presented in section 4.3.4 does not consider competing risks. In Chapter 2, I described how such analyses in the presence of competing risks leads to overestimation of the probability of the outcome of interest. In this section I therefore present the results of competing risk analysis with in-hospital death as the outcome of interest and discharge alive as the competing risk. I present these results in three parts

i. Overall
ii. By sex
iii. By birthweight categories

These analyses are presented as cumulative incidence curves which show the absolute probability of in-hospital death and discharge alive at different time points following admission to the neonatal unit.

*a) Overall*

Data were available for 9074 out of the 9115 subjects admitted. In-hospital death was observed in 840 neonates while 8126 were discharged alive. The 108 referrals were right censored. The total duration of follow up was 98 days. Figure 4.10 shows the absolute probabilities of in-hospital death and discharge alive over time.

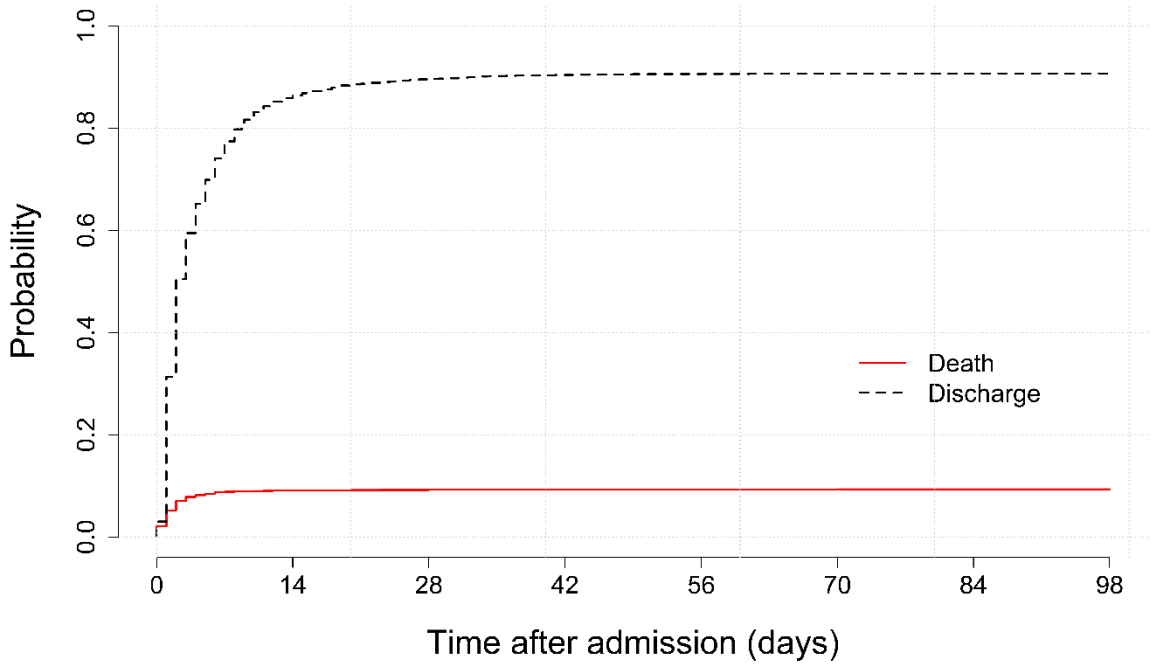



***Figure 4.10**: Probability of in-hospital death or discharge alive for all neonates*

The probability of being discharged alive rapidly rises and remains higher than the probability of in-hospital death throughout follow up. Probability of in-hospital death peaks out at less than 10% which is the overall case fatality rate of 9.2% observed in Table 4.1. This peak is attained within the first week after admission, with 22% (188/840, 95% C.I 20 to 25%) and 73% (613/840,95% C.I 70 to 76%) of these deaths occurring in the first 24 hours and between 24 hours to 7 days respectively.

*b) By sex*

A total of 9033/9115 observations were available to analyse competing risk survival by sex. In this sample, there were 840 in-hospital deaths and 8085 neonates were discharged alive overall. Sample size by sex were: males 449 dead, 4441 alive (CFR 449/4944, 9.1%) and females 391 dead, 3644 alive (CFR 391/4089,9.6%). There were 108 referrals in total and these observations were right censored. The competing risk survival curves by sex are illustrated in Figure 4.11

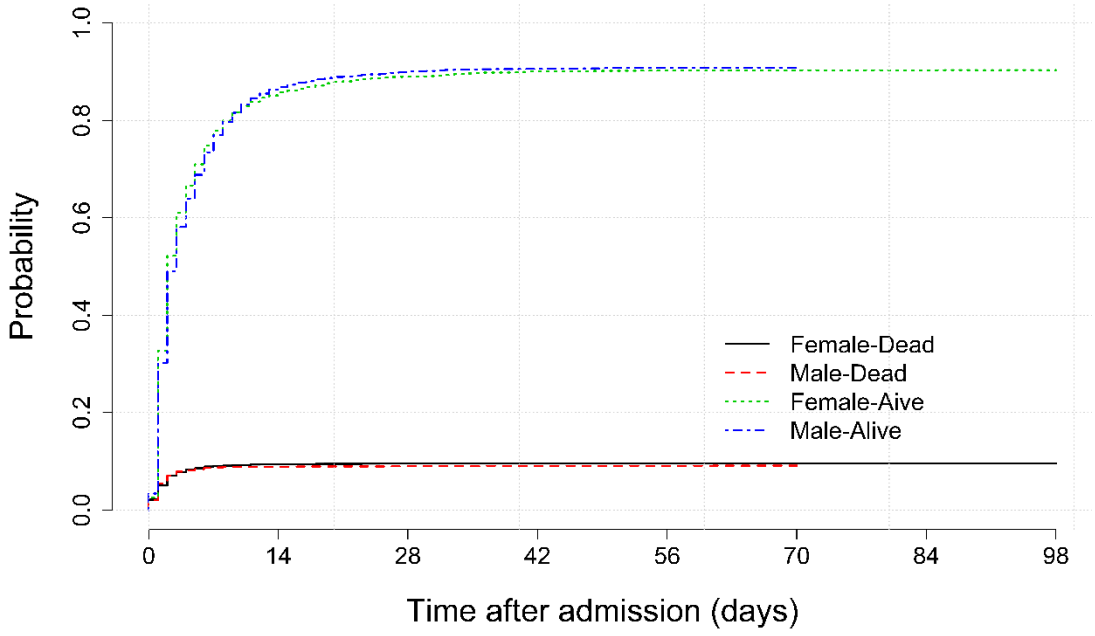



***Figure 4.11**: Probability of in-hospital death or discharge alive by sex*

Figure 4.11 demonstrates that there was no difference in probability of in-hospitality death and probability of being discharged alive by sex. The probability of in-hospital death for both sexes peaks out at 9.2% which is the same as the crude mortality rate for the study population. The last observations are however at different time points with males at day 70 and females at day 98.

*c) By birthweight*

In the 2 preceding sections (competing survival analysis for the overall population and by sex), results are presented as a single graph where probability of discharge alive is stacked on top of probability of in-hospital death. In this section, similar graphs are presented but each deals with one of the five birthweight categories (Chapter 1, Table 1.6): ELBW (<1kg), VLBW (1-<1.5kg), LBW (1.5-<2.5kg), Normal (2.5-4kg) and Macrosomia (<4kg).

I begin with the two lowest categories i.e. ELBW and VLBW in Figure 4.12. For ELBW group, there were 49 observations (45 dead, 1 alive, 3 referred) with CFR of 92% (45/49). For the VLBW, there were 280 observations (192 dead, 79 alive, 9 referred) and a CFR of 69% (192/280).

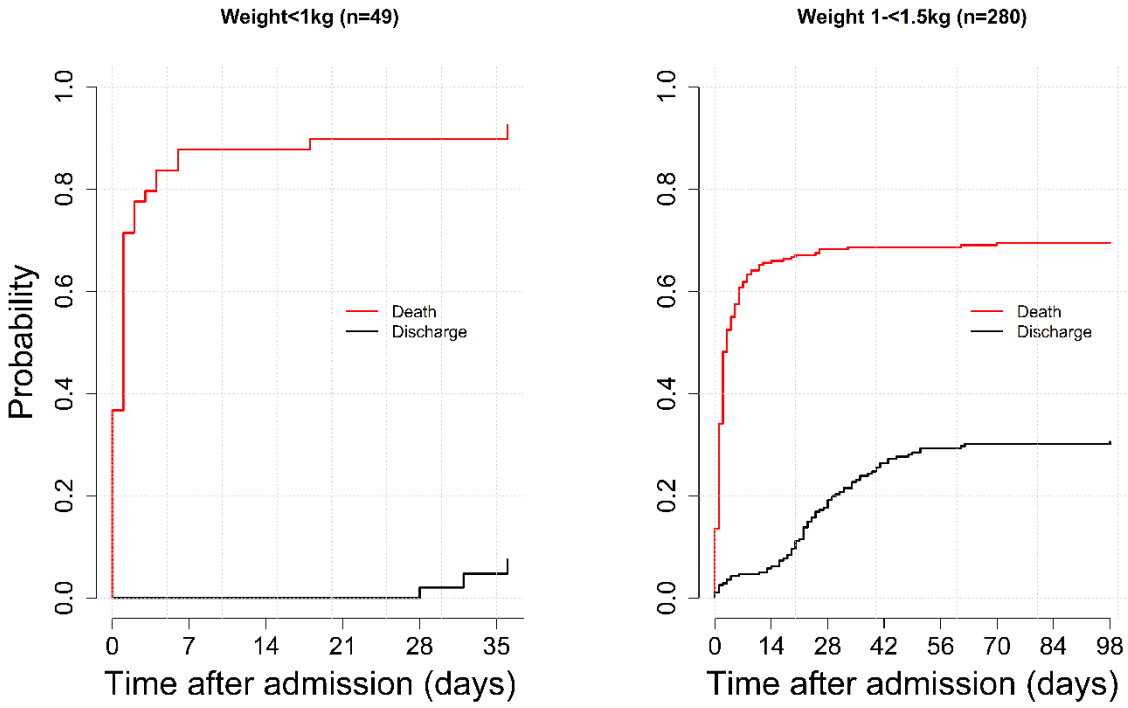



***Figure 4.12**: Probability of in-hospital death and discharge alive for birthweight less than 1.5kg*

Figure 4.12 contains two graphs showing the probability of in-hospital death and discharge alive over time for ELBW and VLBW neonates. There was a difference in the maximum length of stay; 35 days vs 98 days for ELBW and VLBW respectively. There was a steep rise in probability of death in the first day and first week of admission, peaking at >0.9 for ELBW and 0.7 for VLBW. In addition, the probability of death remains higher than the probability of discharge alive throughout the course of hospital stay for both weight categories.

There was no discharge alive observed in the first 28 days after admission for ELBW, while there was a gradual rise in the first 2 weeks for VLBW. The steepest rise in the probability of being discharged alive for VLBW neonates was seen between 2 weeks (14 days) and 8 weeks (56 days) post admission. These

difference in patterns of in-hospital deaths and being discharged alive results in the wider gap between the two cumulative incidence curves seen in ELBW as compared to the VLBW.

The next graph (Figure 4.13) depicts survival of neonates in the LBW category. The definition of this category -1.5 to less than 2.5 kg- is widely used in neonatal medicine. I present this category separately as it's the widest category amongst the three low birth weight categories with a 1 kg interval. While ELBW is < 1kg the lower bound is effectively 500g as this is considered a threshold of viability. There were a total of 2120 observations (257 dead, 1828 alive, 35 referred) with CFR of 12%(257/2118).

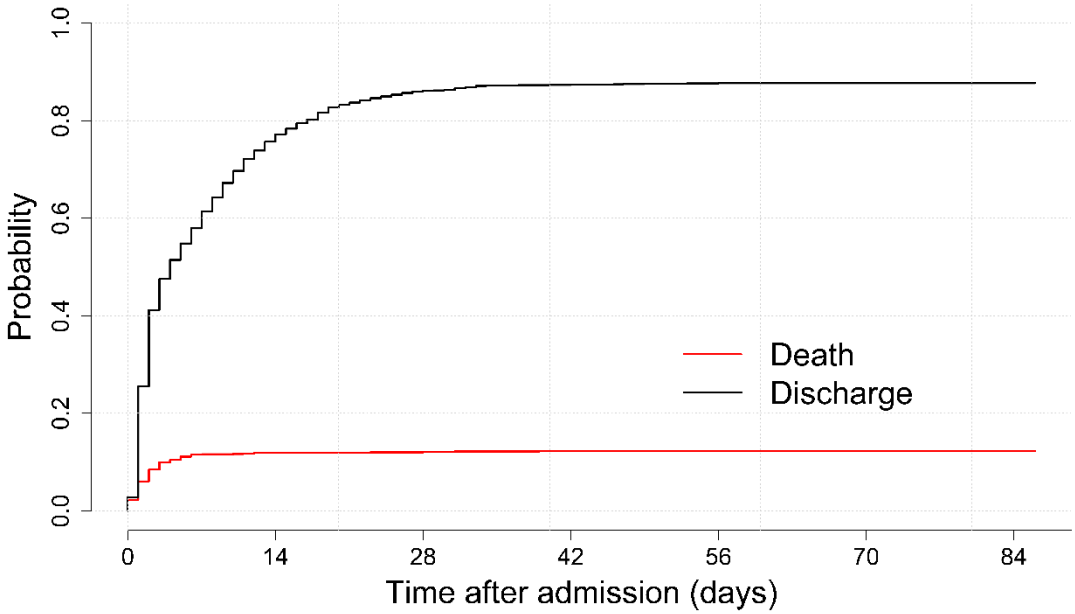



***Figure 4.13**: Probability of in-hospital death or discharge alive for birthweight 1.5 to 2.499 kg*

A key feature in Figure 4.13 is that the cumulative incidence curve for the probability of discharge alive is above the probability of in-hospital mortality in contrast to Figure 4.12. This shows that the probability of discharge alive is greater than probability of in-hospital death. However, note the closeness of the curves in the first days after admission. Considering these two observations, I went ahead to investigate further this switch in probability of outcome by splitting the category into two 500 g groups i.e. 1500-1999g and 2000-2499g. The result is illustrated in Figure 4.14.

The two graphs in Figure 4.14 includes 670 neonates in the 1500-1999g category (157 dead, 493 alive,20 referred) and 1450 neonates in the 2000-2499 category (100 dead, alive 1335, referred 15). The case fatality ratio is 23% (157/670) vs 7% (100/1448) for 1500-1999g and 2000-2499g categories respectively.

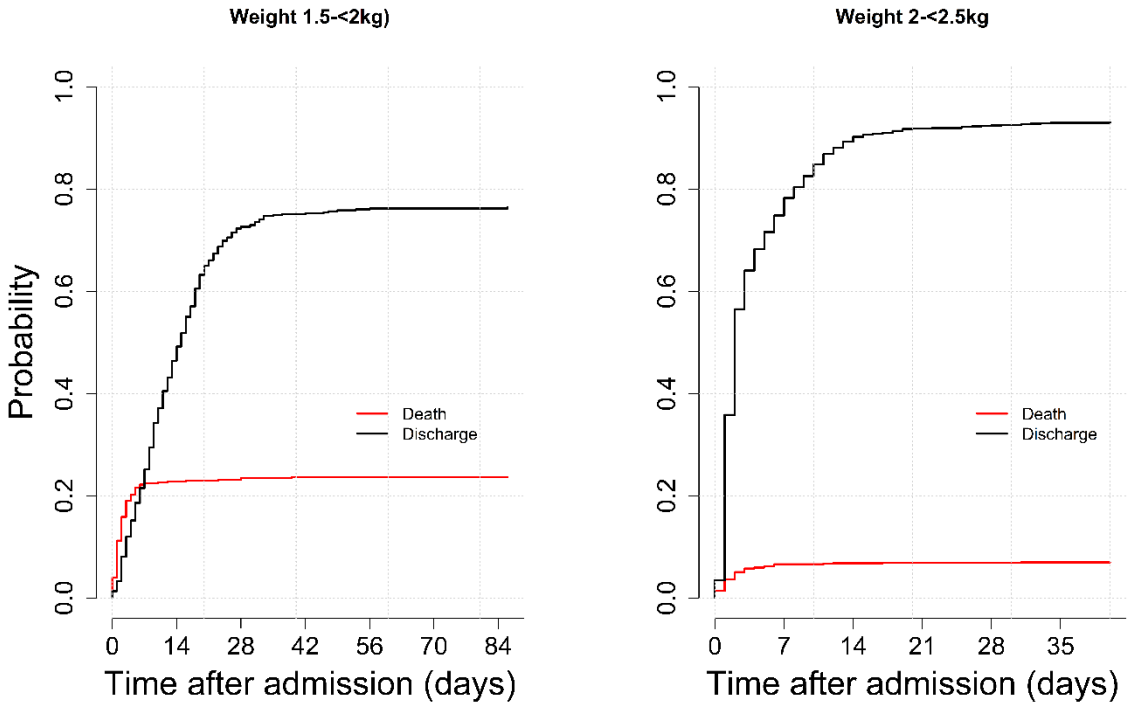



***Figure 4.14**: Probability of in-hospital death or discharge alive for birthweight 1.5 to 1.999 and 2 to 2.499kg*

From Figure 4.14, it is evident that the switch in survival probability occurs in the 1500-1999 category at just about the end of the first week of admission where we see the cumulative incidence curves crossing. After this point, the probability of discharge alive remains higher than probability of in-hospital death at all time points after admission. The peak probability of risk of in-hospital death rises faster and attains a higher peak in the 1500-1999g category compared to 2000-2499; 23% vs 7% respectively.

Finally, I consider the two birthweight categories above 2.5 kg, normal weight (2.5-4kg) and macrosomia (greater than 4kg) together in Figure 4.15. There were 6253 neonates in the 2.5 to 4kg category (336 dead, 5861 alive, 61 referred) with CFR

of 5% (336/6253). The macrosomia category had 352 (7 dead, 343 alive, 2 referred) and a CFR of 2% (7/353).

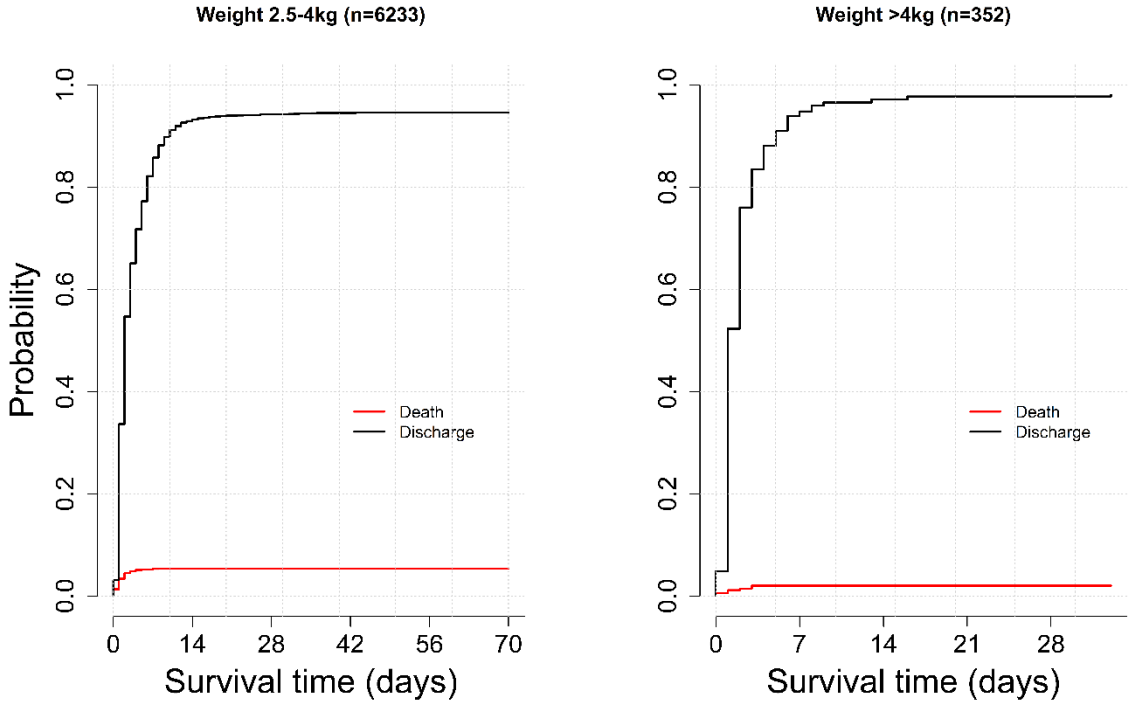



***Figure 4.15**: Probabilty of in-hospital death or discharge alive for birthweight 2.5 to 4k and greater than 4 kg*

Figure 4.15 shows that the maximum length of stay observed in normal weight babies was more than twice that observed in those in the macrosomia category, 70 days vs 33 days respectively. The lowest estimated absolute probability of in-hospital death is observed in these two categories at less than 0.05.

## 4.4 Discussion

### *4.4.1 Summary of findings*

In this Chapter, four questions were answered to meet the first objective of the DPhil to describe; clinical characteristics of the study population, case fatality rate, length of stay and time to in-hospital death. During the study period spanning a total of 18 months, 9115 neonates were admitted to the neonatal unit. There was variation in the number of admissions observed per month: median 595, range 531 to 685 in 2014 and median 406, range 273 to 613 in 2015. Variation is also observed in daily admissions: median 20, range 4 to 35 in 2014 and median 13, range 2 to 30 in 2015.

Overall about 1 in 10 (9%) of neonates admitted died with the highest case fatality among the ELBW (92%) and the lowest among those with Macrosomia (2%). Whereas the risk of in-hospital death among the normal birth weight neonates (5%) was about half the overall risk, the absolute number of deaths (336/840 or 40% of all deaths) was the highest as the majority (70%) of the admissions were of this category. The overall median length of stay was 2 days with a range of 0-98 days. The Kaplan-Meier analyses of time to in-hospital death graphically demonstrated that the ELBW neonates experienced short stays and the highest probability of death in contrast to macrosomic neonates who had short stays but with the lowest probability of death. Finally, the competing risk survival analyses showed that the probability of in-hospital death is higher than that of discharge alive in ELBW and VLBW categories. In addition, splitting the LBW category further revealed that the probability of discharge rises to be higher than that of death in the 2 to less than 2.5kg sub-category.

*4.4.2 Case fatality rate*

The overall case fatality rate (CFR) in this study (9.2%) lies within the range observed in other Kenyan hospitals shown on Table 1.8 (3%- 62%). However, the extremes of this range represent special populations i.e. the lowest (3%) was observed in neonatal intensive care in a tertiary private hospital(79) while the highest (62%) included infants with neonatal tetanus only in a rural district hospital(81). Nonetheless, 9.2% is low relative to the majority (12/15) of the other Kenyan hospitals which reported CFR >10% (Table 1.8). Similarly, it is in the lower end of the range of 3-35% reported in other African countries (Table 4.1) given that most articles (18/23) reported CFR of 10% or greater. However, the majority (16/23) of these African studies were conducted in tertiary hospitals rather than in a district hospital level facility.

The CFR in VLBW neonates was 69%, a figure much higher than 45% (75) and 44% (73) in a Kenyan tertiary hospital, 48% in Malawi (238) and 38.9% in Tanzania.(238) This CFR is however similar to the 68.9% reported from a rural NICU in Uganda.(237) This high risk of death, reflects the known association between very low birth weight and mortality (52) and more so among the VLBW who disproportionately contribute to neonatal mortality (256).This highlights how an overall CFR including all birthweight categories masks significant differences by weight category.

Data from high income settings reveals much lower risk of in-hospital death following admission to a neonatal unit. The Australia and New Zealand Neonatal Network includes all NICUs in the two countries plus one from Singapore. In

addition, there also 16 special care units (a level lower than NICU) in Australia and 14 in New Zealand contributing data.(257). In their 2015 report, out of 9672 from NICU ,95.5% survived to discharge while 98.6% of 1219 in special care units survived.(258) This is comparable to survival in the Canadian Neonatal Network (CNN) where in 2015, out of 14,807 neonates in 30 participating facilities 97.5% survived to discharge.(259) In data for 408,164 VLBW and ELBW neonates (501-1500g) from facilities participating in the Vermont Oxford Network in the United states, over a ten year period (January 1, 2005, to December 31, 2014), rate of death decreased from 14.0% in 2005 to 10.9% in 2014. (260) This is in marked contrast to the 70% for VLBW and 92% for ELBW observed in this study.

The overall CFR in the study facility is therefore relatively low compared to Kenyan and African neonatal units but high compared to CFR in high income countries. The CFR for VLBW and ELBW neonates is however markedly high. Such variation in hospital mortality is a well recognised phenomenon arising from multiple complex factors. It is however difficult and indeed incorrect to make direct comparisons of mortality between hospitals without any effort to adjust for differences in case mix. One way to adjust for case mix is through the use of prognostic models such as the two models described in Chapter 5 and 6. Understanding mortality within hospitals nevertheless remains important particularly for understanding and improving the quality of care in addition to tracking change over time.(221, 222) This may be applied within clinical networks which are a useful strategy for improving patient care.(257)

*4.4.3 Length of stay*

The distribution of length of stay in the study period was right (positively) skewed. This is a well recognised characteristic of length of hospital stay in all age groups.(261) Most of the observations with regards to length of stay (in survivors and deaths) were clustered around the first week and especially over the first two days after admission. A much smaller number of neonates exited the unit after a week (1216/9107, 13%) stretching up to more than 3 months (98 days). This relatively small number with stay greater than 1 week are however likely to account for a disproportionately big part of resources utilised in the neonatal unit.

There are few data to compare with from Kenya with only 1 out of the 15 identified articles in Table 1.8 reporting length of stay.(251) In this article, Myhre et al reported mean and range in days with 39 (10–112) vs 28 (16–54) before and after introduction of CPAP respectively.(251) This paucity of data extends to the rest of the African studies described in Table 4.1 . Given the skewed nature of these data, out of the 7 articles that reported length of stay (231, 235, 237, 242, 246, 247, 249), only two reported median and range. i.e. Shah and Zash (242, 246). In Pepler et al the median length of stay was 11 days (range 1-171) but this included data on those who were discharged alive only.(246) Zash and colleagues found a median of 5 days (interquartile range 2-15).(242)

A systematic review of factors affecting length of stay in neonatal units identified birthweight amongst the key factors alongside gestational age and sex.(262) In high income settings, low in-hospital mortality has resulted in longer hospital stays with an attendant increase in resource use.(262) The Australia and New Zealand

Neonatal Network is a case in point where the median length of stay for babies who survived NICU care until discharge home was 27 days (interquartile range 10-52).(258) The authors of the report however noted that length of stay increased with a decrease in gestational age. The most premature (<24 weeks) stayed the longest, median 131.5 days (interquartile range 119.5-151.5). This is in sharp contrast to my study where taking birthweight as a proxy for gestation (in the absence of gestational age data), the smallest (ELBW) babies stayed for a very short time median 1 day (interquartile range 1-2 days) due to extremely high mortality (92%).

If neonatal units in low resource settings achieve the anticipated improvements in survival of the admitted neonates, it is anticipated that there will be a concomitant increase in length of stay and resource use. As an example, if the CFR rate for VLBW in the study facility is reduced from 70% to the 11% seen in the American Vermont Oxford Network(260) the number surviving to discharge would increase from 79/280 (Figure 4.4) to 249/280. Assuming the same median length of stay for VLBW who survive to discharge (23 days, Figure 4.4), it would mean the facility would then take care of an extra 170 VLBW for a median 23 days. This would lead to a considerable increase in costs and therefore should be a key consideration in service delivery planning and organisation in low resource settings as hospital based care for neonates is developed further.

*4.4.4 Time to in-hospital death*

*a) Kaplan-Meier*

In this study I analysed the time to in-hospital death using both Kaplan-Meier and competing risk approaches. The key finding from the Kaplan-Meier analyses was the demonstrated difference in survival by birthweight ($p<0.001$, Log rank test). None of the Kenyan studies described in Table 1.8 included any time to event analyses. In the reports from other African countries on Table 4.1, two studies included Kaplan-Meier analyses but none of these examined survival by birthweight. Tochie et al found better survival for neonates without Neonatal Respiratory Distress Syndrome compared to those with ($p<0.0001$ for both Wilcoxon and Log rank test).(234) Zash et al compared mortality by HIV exposure status and found no difference ($p=0.19$) but demonstrated a poorer survival curve in babies who were not fed enterally (but provided with intravenous fluids) compared to those who were fed enterally (no Log rank test reported).(242) In the presence of competing risks (in this case in-hospital death vs discharge alive) Kaplan-Meier analysis assumes a hypothetical situation where in-hospital death occurs in a setting where no patients are discharged alive.(263) This is clearly unrealistic and I discuss the competing risk analyses in the next section.

*b) Competing risks survival analysis*

Figure 4.16 compares the overall cumulative incidence of in-hospital death (derived from the complement of Kaplan-Meier estimates in Figure 4.7) with estimates using competing risk analysis (derived from Figure 4.10)

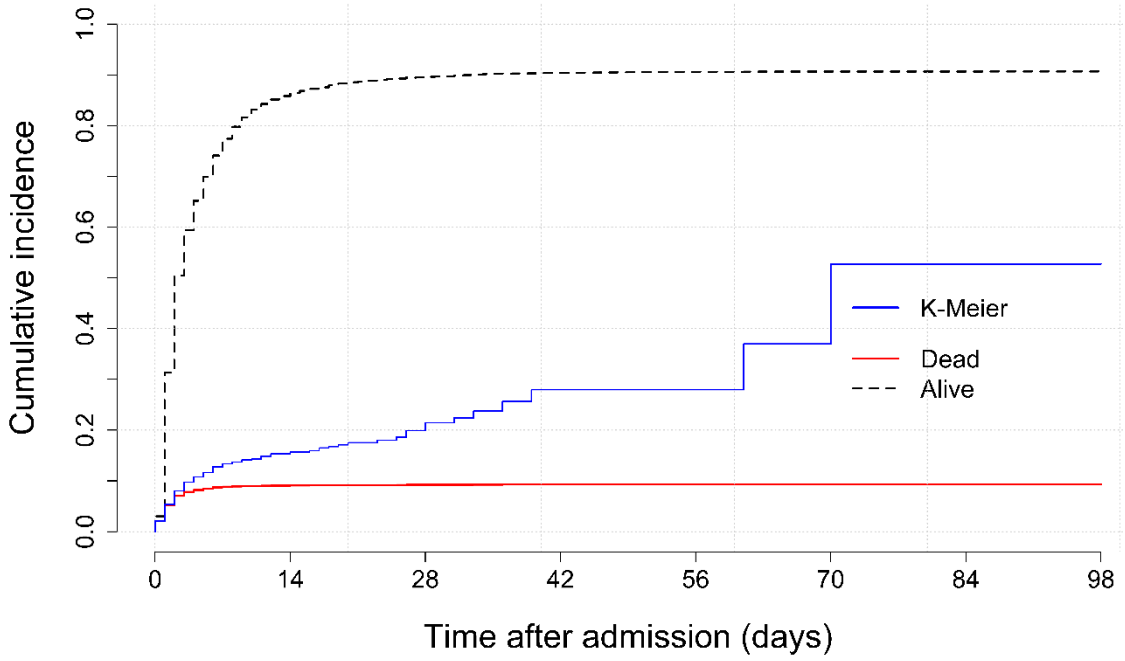



***Figure 4.16**: Comparison of cumulative incidence of in-hospital mortality by Kaplan-Meier and Competing Risks methods*

At each time point in Figure 4.16, the cumulative incidence of in-hospital death by Kaplan-Meier greater is greater than that by the competing risks approach. As a result, the estimated cumulative incidence of in-hospital death over the period of observation by Kaplan-Meier is 53% in contrast to 9.4% by competing risks analysis. From Table 4.2, the crude case fatality ratio in the study population was 9.2% and therefore the Kaplan-Meier method in this case overestimates cumulative incidence of in-hospital death.

This finding is due to bias caused by the reduction in the risk set as a result of right censoring of patients who were discharged alive (Paragraph 4.2.9 and Chapter 2,Paragraph 2.5.3).(263) This shows that the complement of the Kaplan-Meier is an inappropriate estimate of crude incidence of in-hospital mortality given the

presence of discharge alive as a competing risk.(263) This competing risk bias has indeed been observed in a review of 100 publications with K-M analyses by Walraven and colleagues. They found that almost half (46%) of these studies contained K-M analyses susceptible to competing risk bias. (264)

Competing risks analysis of in-hospital death and discharge alive by birthweight category demonstrated that the probability of dying in the neonatal unit is higher than that of being discharged alive for the two lowest weight categories (ELBW and VLBW, Figure 4.9). A switch in survival probability is subsequently observed in the LBW category where for infants weighing 2000-2499g, the probability of discharge alive rises above that of in-hospital death (Figure 4.11). This difference within the LBW category suggests that it may be useful at the very least to have a “high” LBW (2000-2499g) and “low” LBW (1500-1999g) category in reports on hospital care for neonates from LMICs. Reporting by more granular weight categories may even be extended to <500 g intervals (e.g. 100g, 250g) in the manner adopted in reports from neonatal networks in high income countries.(257, 259) Nevertheless, for all the five birth weight categories, most of the deaths occur in the first week following admission.

There were no other instances of competing risk analysis in neonatal care pertaining to hospitals in Kenya or other African countries identified in the recent literature (2007-2017, Table 1.8 and Table 4.1). In an example of the application of the competing risk method to neonatal care in a high income setting, Hinchliffe et all computed the probability of dying, being discharged and being cared for in the unit over time. This was done using data on neonates with a gestational age of

between 24-28 weeks admitted to 29 neonatal units over a 5-year period (January 2006 to December 2010).(252) A key difference between this study and this thesis is that the former was restricted to extremely preterm (<28 weeks gestation) babies. However, two key similarities are observed. Firstly, the highest rates of in-hospital death occur in the initial weeks after admission. Secondly, the probability of survival to discharge rose with increase in birthweight.(252)

The competing risk analyses as demonstrated in this thesis could therefore be useful in communicating the probable length of stay by birthweight category to clinicians caring for neonates in high mortality, low resource settings. Such information would be useful in characterising variation between units, informing resource allocation and for communicating to parents.

With regards to variation between units, the simultaneous estimation of probability of in-hospital death and discharge alive at different time points by clinical characteristics e.g. birthweight would allow a more detailed comparison of patients seen in different facilities than that afforded by overall case fatality rates. Hospital costs are closely linked to length of hospital stay and data on when patients are likely to leave the unit are thus useful for projecting resource requirements and by extension inform resource allocation. This is particularly important in low resource settings hospitals where multiple needs compete for limited resources. With the anticipated improvement in survival to discharge particularly for the VLBW and ELBW, the resultant long hospital stays mean accurate estimates of probability of discharge at different time points - as provided by competing risk analyses - will be useful in counselling parents on when they are likely to leave for home

*4.4.5 Categorisation of birthweight for average and individual prognosis*

In this Chapter, I have described case fatality, length of stay and time to in-hospital death by birthweight categories. These analyses describe the average prognosis in the study population.(68) Categorisation of continuous variables such as birthweight may lead to loss of statistical power, concealment of non-linearity and residual variation within categories.(265) I have indeed demonstrated residual variation in the LBW category (Section 4.3.5 (c), Figure 4.14). In describing average prognosis, this phenomenon may be dealt with by either retaining birthweight as a continuous variable or as noted in section 4.4.4 (b) use of smaller birthweight categories e.g. 500g intervals.

However, when estimating absolute risk of in-hospital death for individual patients (prognostic modelling), retaining the birthweight categories typically used in clinical practice is advantageous due to clinicians' familiarity with these. In Chapters 5 & 6, I deal with this residual variation in *average prognosis* by birthweight categories by using additional clinical characteristics to develop multivariable prognostic models that estimate the probability of in-hospital mortality at *individual patient level*. (70)

*4.4.6 Limitations*

These data are from one site and therefore represent only one context within a low resource setting. This facility has the capacity to offer more advanced respiratory support (BCPAP) albeit with only 2 machines which is unusual in public sector neonatal units in Kenya. However, the facility has in common with other low

resource settings the provision of essential neonatal care with resource constraints particularly low nurse to patient ratios

Gestational age is a key risk factor for neonatal outcomes and I could not incorporate this into my analyses due to missing data. However, good availability of birthweight provided invaluable insights. This does not diminish the need for data on gestational age. It is a well recognised determinant of survival with prematurity (gestation less than 37 weeks) associated with increased risk of postnatal mortality and morbidity.(266) In addition, accurate gestational age is required in diagnosing (alongside weight) intrauterine growth restriction (IUGR) which leads to small for gestational age (SGA) babies. SGA is defined as infants below the 10th centile of a birthweight-for-gestational-age, gender-specific reference population.(267) This suboptimal growth is an important cause of mortality and morbidity.(268) Lack of gestational age data such as demonstrated in this study contributes to failure to recognise this important condition in LMICs where the burden is estimated to be very high (32.4 million in 2010).(267)

In-hospital mortality may be subject to bias as there may be variation in discharge criteria particularly when discharge is followed by unobserved early post discharge mortality. The absolute risks of in-hospital death presented in this Chapter apply to the average patient in the study population and it is expected that there would be residual variation at individual patient level. However, these data are useful for informing decision making particularly for groups of patients and for service planning as described in Table 1.5 of Chapter 1.

### *4.4.7 Conclusion*

Case fatality, length of stay and time to in-hospital death are important outcomes with respect to fundamental prognosis in neonatal hospital care. The observed difference in length of stay for the ELBW and VLBW overall and those discharged alive suggests that improved survival for these neonates will lead to longer hospital stays with attendant increase in costs requiring planning for this scenario. Attention also needs to be paid to the normal weight babies with regards to reduction of absolute numbers of deaths. The cumulative incidence function provided a better estimate of cumulative incidence of in-hospital mortality and demonstrated residual variation in risk of death. This residual variation means better methods of estimating risk of in-hospital mortality particularly at individual patient level such as prognostic modelling are required.

Data from this study are likely to be applicable to other district hospital level neonatal units in LMICs where intensive care is not available as evidenced by similarities in the case fatality rate. Nevertheless, more of such data are required from other neonatal units in Kenya and other low resources settings (akin to neonatal clinical networks in high income countries such as the Canadian Neonatal Network) to support service improvement and monitoring especially as more advanced neonatal care is scaled up.(259)

# Chapter 5 : Development and validation of the Neonatal Essential Treatment Score (NETS)

## 5.1: Introduction

### *5.1.1 Background*

In Chapter 4, I described crude case-fatality ratios and time to in-hospital death as measures of average prognosis for the study population.(68) Prognostic modelling provides a means of extending description of prognosis from the population to the individual patient level. This entails the combination of multiple predictors in a model to estimate the risk of the outcome of interest occurring within a specific time period for individual patients.(269) In this case, my specific interest was the prediction of in-hospital mortality. As outlined in Chapter 1 and Chapter 3, the predicted risk of in-hospital death may be useful in clinical decision making, case mix adjustment and service delivery planning thus contributing to efforts to improve neonatal survival through hospital care.

In Chapter 3, I presented a systematic review of prediction models that have included therapeutic interventions as predictors (treatment intensity scores) in neonatal medicine. The review concluded low certainty in the evidence of the performance of these models in predicting in-hospital mortality, morbidity and resource use in low resource settings. There were two principal limitations identified in the included studies. First, the therapeutic interventions used are typically given in a NICU setting and thus many are not applicable as predictors in low resource hospitals without NICU care. Secondly, I identified numerous limitations in the methods used to develop and validate these models.

In this chapter, I present the first of two models I developed to predict in-hospital mortality using routinely available data. The second model is presented in Chapter 6. This chapter therefore, addresses part b of objective 2 of the thesis; predicting in-hospital mortality using a treatment intensity scoring approach. This will be referred to as the Neonatal Essential Treatment Score (NETS). In Section 5.1.2, I describe the selection of therapeutic interventions that will be used as predictors in settings offering essential neonatal care. In Section 5.1.3, I describe the methodological steps I undertook for developing NETS to overcome the limitations I identified in the review.

*5.1.2 Essential neonatal treatments as predictors of in-hospital mortality*

*a) Selection of therapeutic interventions for in-patient essential newborn care*

Essential Newborn Care was defined in Chapter 1 as "a comprehensive strategy designed to improve the health of newborns through interventions before conception, during pregnancy, and soon after birth, and in the postnatal period". This term is used specifically in reference to LMICs to focus resources on effective low cost interventions.(12) The second objective of this thesis is to build statistical models to predict in-hospital mortality following admission to the neonatal unit. Therapeutic interventions that might be predictors were thus restricted to elements of essential newborn care provided in the postnatal period after admission to the neonatal unit. This potentially increases the clinical utility and generalisability of the model to LMIC neonatal units providing essential newborn care.

I derived candidate predictors (interventions) from clinical practice guidelines for in-hospital post-natal care for neonates in low resource settings developed by the

WHO and locally adapted for Kenya.(14, 17, 270) These are described in Table 5.1.

***Table 5.1****:Therapeutic interventions used for in-patient, post-natal essential newborn care*

| Problem | Intervention (s) |
|---|---|
| Respiratory support | Oxygen, BCPAP* |
| Circulatory support if feeding is contraindicated | Intravenous fluid infusion |
| Feeding where breastfeeding not possible | Enteral feeds (preferable expressed breastmilk) |
| Possible severe bacterial infection | Penicillin (or ampicillin) and gentamicin |
| Hypothermia | Kangaroo Mother Care |
| Neonatal seizures | Glucose, Phenobarbital |
| Neonatal jaundice | Phototherapy/exchange transfusion |

*Bubble Continuous Positive Airway Pressure

These interventions address the major causes of neonatal mortality in low and middle income countries: complications of preterm birth, intrapartum related complications (including birth asphyxia) and neonatal infections.(4) These predictors were further reduced based on clinical knowledge and described in Section 5.2 where I present the final list.(136)

*b) Addressing the challenges of using treatments as predictors*

Therapeutic interventions typically pose a challenge to the predictive accuracy of prognostic models if the treatments reduce the risk of the predicted outcome. This alters the relationship between predictors and outcomes thereby resulting in incorrect predictions when the model is used in a population that is either untreated or is treated differently.(271) This effect of treatments is greatest for therapy provided after predictions have already been made (i.e. during follow up). The model I present in this chapter however uses treatments prescribed at baseline before computation of the predictions. Approaches to handle treatments in

prognostic modelling include either ignoring them altogether or including them as predictors. The latter approach is preferable based on empiric evidence.(271)

The therapeutic intensity approach explicitly includes treatments as predictors but faces potential drawbacks. The first, as described in Chapter 3 is variation in clinical practice which would result in different treatments but this is mitigated against in the Kenyan context by the relatively high degree of standardisation of essential neonatal care in the County hospitals.(17) In addition, treatments may change during the course of hospital stay and for this reason I only considered treatments prescribed at the time of admission (the time point of intended use of NETS) particularly as most neonatal mortality is observed in the first 24 hours of life.(1, 224)

*5.1.3 Rationale*

Predicting in-hospital mortality can provide useful prognostic information for decision making to improve in-hospital care and thereby survival for neonates in LMICs as described in Chapter 1 (Section 1.6) and Chapter 3 (Section 3.1). The systematic review in Chapter 3 suggested the potential of treatments in predicting in-hospital neonatal mortality. Key considerations in predictor selection include availability, low cost of measuring the predictors and precision of measurement (for accurate data with low levels of misclassification).(136) There is evidence that data on treatments prescribed at the time of admission in Kenya meets the availability criterion and the data collection system for the study data set as described in Chapter 2 (Section 2.3) addresses accuracy.(31, 32) I therefore applied recommendations for the development, validation and reporting of prognostic

models as described in Chapter 2 to develop a treatment intensity scoring approach for the prediction of in-hospital mortality to suit Kenya and potentially other LMICs settings.(92, 135, 136, 269, 272) These include the use of bootstrapping to evaluate for overfitting (internal validation), emphasis on external validation as the most important measure of predictive accuracy and use of the TRIPOD statement for reporting the study.

### *5.1.4 Objective*

I aimed to use essential neonatal treatments prescribed at the time of admission to a large urban neonatal unit as predictors to:

a) Develop and internally validate a neonatal treatment intensity score to predict in-hospital mortality
b) Externally validate the model in a data set obtained from the same unit from a different time point (temporal external validation)

## 5.2 Methods

In Chapter 1 (Figure 1.8), I outlined the four stages of prognostic model development: development (derivation), validation, impact assessment and implementation.(95) Thereafter in Chapter 2 (Section 2.6), I further described the general principles underpinning the two steps undertaken in this thesis, development and validation. In this section, I will now describe the specific application of these principles (development and validation) in the development of the NETS model. In doing this I considered six of the seven steps for developing and validating prediction models proposed by Steyerberg and Vergouwe: (1) problem definition and data inspection, (2) coding of predictors, (3) model specification, (4) model estimation, (5) model performance and finally (6) validation (internal and external).(135) This description for the most part also applies to the Score for Essential Neonatal Symptoms and Signs (SENSS) which is described in Chapter 6.

### *5.2.1 Derivation*

*a) Source of data*

The dataset used to derive the model was obtained from a routine data collection system set up in the study neonatal unit which I described in detail in Chapter 2 (Section 2.3). Follow up was calculated in days from time of admission to time to exit from the unit which followed either in-hospital death, discharge alive or referral. The outcome of interest for the NETS model is in-hospital death.

*b) Participants*

The study participants for model derivation were all the patients admitted to the study neonatal unit between April 2014 to December 2015 while those for external validation spanned all admissions from January to July 2016. All the patients were entered into the data base consecutively at the point of discharge. Patient flow for the derivation data set is summarised in Figure 5.1.

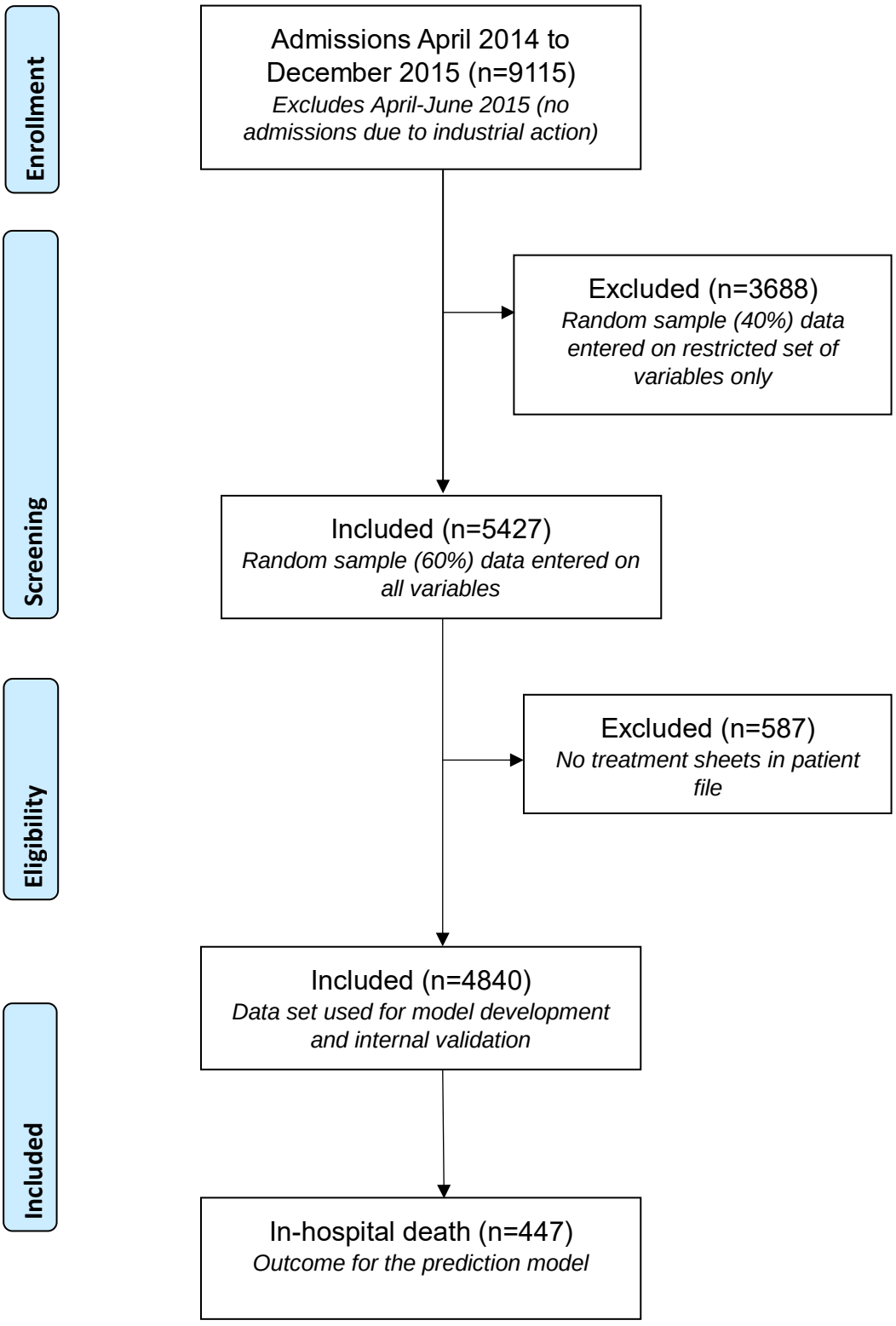



***Figure 5.1**: Patient flow chart for NETS model derivation*

A total of 9115 patients were admitted during the study period for the derivation data set. 5427 patients were randomly selected for data entry on the full set of variables as described in Chapter 2 (which included treatments prescribed at admission) and were therefore eligible for inclusion. I excluded the 3688 who had data on the restricted set of variables only (the restricted set of variables did not include treatments amongst other variables). Out of the 5427 eligible patients, 587 were missing treatment sheets in their case records which I excluded leaving 4840 patients to develop the NETS model. In-hospital death occurred in 447(9.2%) of these patients.

*c) Outcome for the prognostic model*

The outcome of interest was in-hospital mortality defined as death from any cause (all-cause mortality) that occurred in the neonatal unit. As described in Chapter 2, data were entered after the patient was discharged from the unit and thus the data collectors were not blind to predictor status. They were however unaware of these analyses conducted only after the period of data collection. In addition, mortality is a hard outcome which is not subject to measurement error and bias.

*d) Predictors*

The main predictors considered were the essential neonatal treatments described in Table 5.1. Exploratory descriptive analysis however revealed that phototherapy, exchange transfusion, and Kangaroo Mother Care were rarely prescribed (<1.5%) at the time of admission so these were dropped. This is clinically expected and therefore likely in other neonatal units. I therefore restricted treatments to five

essential treatments prescribed at the time of admission and these are described in Table 5.2.

***Table 5.2****: Essential treatments prescribed at admission that were included as predictors*

| Variable name | Description | Definition | |
|---|---|---|---|
| | | Yes | No* |
| **Intravenous antibiotics** | Penicillin AND gentamicin are first line antibiotics prescribed together for possible Severe Bacterial Infection. | Prescribed | Not prescribed |
| **Intravenous fluids** | Any prescription of an infusion of intravenous fluid e.g. 10% Dextrose | Prescribed | Not prescribed |
| **Enteral feeding** | Expressed breast milk or formula fed through nasogastric tube or cup | Prescribed | Not prescribed |
| **Oxygen** | Supplemental oxygen provided by any method e.g. nasal cannula | Prescribed | Not prescribed |
| **Phenobarbital** | Prescription of phenobarbital | Prescribed | Not prescribed |

*****Reference category in the model

In addition to the treatments, I also included birthweight category and sex as additional predictors in the model as defined in Table 5.3. Birthweight was included since we observed in Chapter 4 that survival in this population varies by birthweight. In addition, while there was no difference in survival by sex in the study population, it is well recognised that there is higher mortality amongst males up to the age of five years at population level.(273)

***Table 5.3**: Additional predictors included in the treatment intensity model*

| Variable | Description | Definition | |
|---|---|---|---|
| **Birthweight** | Weight at birth measured in kilograms as documented at admission. This is a continuous variable but in clinical practice conventionally categorised into 5 categories | ELBW | <1kg |
| | | VLBW | 1 - 1.499kg |
| | | LBW | 1.5 - 2.499kg |
| | | Normal* | 2.5 - 4kg |
| | | Macrosomia | >4kg |
| **Sex** | Documented gender at the time of admission | Male | |
| | | Female* | |

*Reference category in the model
ELBW: Extremely Low Birth Weight, VLBW: Very Low Birth Weight, LBW: Low Birth Weight

I also considered including gestation at birth but as shown in Chapter 3 and 4 this is poorly documented in routine data. In this data set specifically, there was 70% missingness and therefore I did not include it. This missingness is consistent with previous work in Kenyan hospitals where more than 50% of observations did not have gestational age documented.(32)

*e) Sample size*

All available data were used to develop and validate the prediction model. However, I assessed these for adequacy with regards to sample size using the events per variable (EPV) approach. The EPV is based on the number of parameters being estimated. There were seven predictors (six binary and one categorical) included in the model. Birthweight comprises five categories and therefore required 4 parameters to be calculated. Adding these four to one each for the six binary variables gave a total of ten parameters in the model against 445 deaths giving EPV of 44. This is substantially more than the widely recommended 10 and even the more recently suggested 20.(136, 203, 274)

*f) Missing data*

From a clinical perspective, it is expected that each patient's admission treatment sheet will contain a varying combination of prescriptions for the essential treatments used as predictors (depending on clinical need as determined by the admitting clinician). In the data set therefore, if a treatment was not prescribed this cannot be interpreted as a missing value. However, as shown in Figure 5.1 out of the 5427 observations eligible for inclusion, 587 did not have any treatment sheets collected in the file and were therefore excluded. A comparison of outcomes between those with and without missing treatment sheets is provided in Table 5.4 in the results section. I considered multiple imputation to be inappropriate as it would not be clinically plausible to impute treatments prescribed. This left 4840 observations for analysis. Regression analysis was thereafter done using a complete case analysis, as descriptive analyses (Table 5.4) showed that the number with missing data was small. Using this approach, all observations with missing data were dropped from the regression model.

*g) Statistical analysis methods*

All the predictors are categorical. Six are binary (all five treatments as listed in Table 5.2 and sex) while birthweight is ordinal. As described in Chapter 2 logistic regression modelling was used with maximum likelihood estimation. This was implemented using the "lrm" function of the "rms" package in the R statistical programming language (version 3.3.3).(128, 137)

I implemented a full model approach whereby all the predictors were included in the model without any variable selection procedures. I considered two model

assumptions: linearity and additivity. Linearity was not relevant in this case as there were no continuous variables. With regards to additivity, interaction terms were not included because it is largely perceived that prediction model performance (for regression models) is not necessarily improved by inclusion of interaction terms. Inclusion of these terms could in fact increase model overfitting.(136) However, I took into account the interaction between penicillin and gentamicin (prescribed together as first line antibiotics) by combining these into one variable (intravenous antibiotics).In addition, I considered interaction between sex and treatments clinically unlikely.

Model performance was investigated evaluating overall performance, discrimination and calibration. I placed emphasis on discrimination and calibration as defined in Chapter 2 as these are the most important metrics.(92) Calibration curves were constructed using the val.prob.ci.2 function.(204) The R code for these analyses are included in Appendix 9

*h) Risk groups*

No risk groups were explicitly created. However, I explored the clinical utility of the predicted probability of in-hospital death using decision curve analysis as a first step in creating risk groups. Decision curve analysis measures the clinical utility (the net benefit) of a prediction model across the range of predicted probabilities as described in Chapter 2 (Section 2.7). These results are presented in Chapter 7.

*5.2.2 Internal validation*

In Chapter 2 (Box 2.4), I described the procedures I used to conduct internal validation for quantifying optimism in model performance using the bootstrap method. I used 1000 bootstrap samples to estimate optimism in the NETS model after derivation (Chapter 2, Section 2.6.4). This was done using the "validate" function of the "rms" package in the R statistical programming language and the code is included in Appendix 9.

*5.2.3 External Validation*

*a) Source of data*

Data were obtained using the same routine data collection system as model derivation described in Chapter 2 (Section 2.3). Data were entered consecutively for all patients at the time of exit using an electronic data entry tool with built in functionality for checking data inconsistency and implausible values. Further error checking was done using R scripts to produce the clean data used for analysis.

*b) Participants*

I performed a temporal external validation of the NETS model as described in Chapter 2, Section 2.6.5. The data used spanned the period January 2016 to July 2016.The participant selection is summarised in Figure 5.2. From a total of 2735 patients, 1108 had data on the restricted set of variables only and were excluded. Out of the remaining 1627, 143 had no treatment sheets in the patient files and were also excluded leaving 1443 observations with 137 in-hospital deaths (9.5%) for the external validation data set.

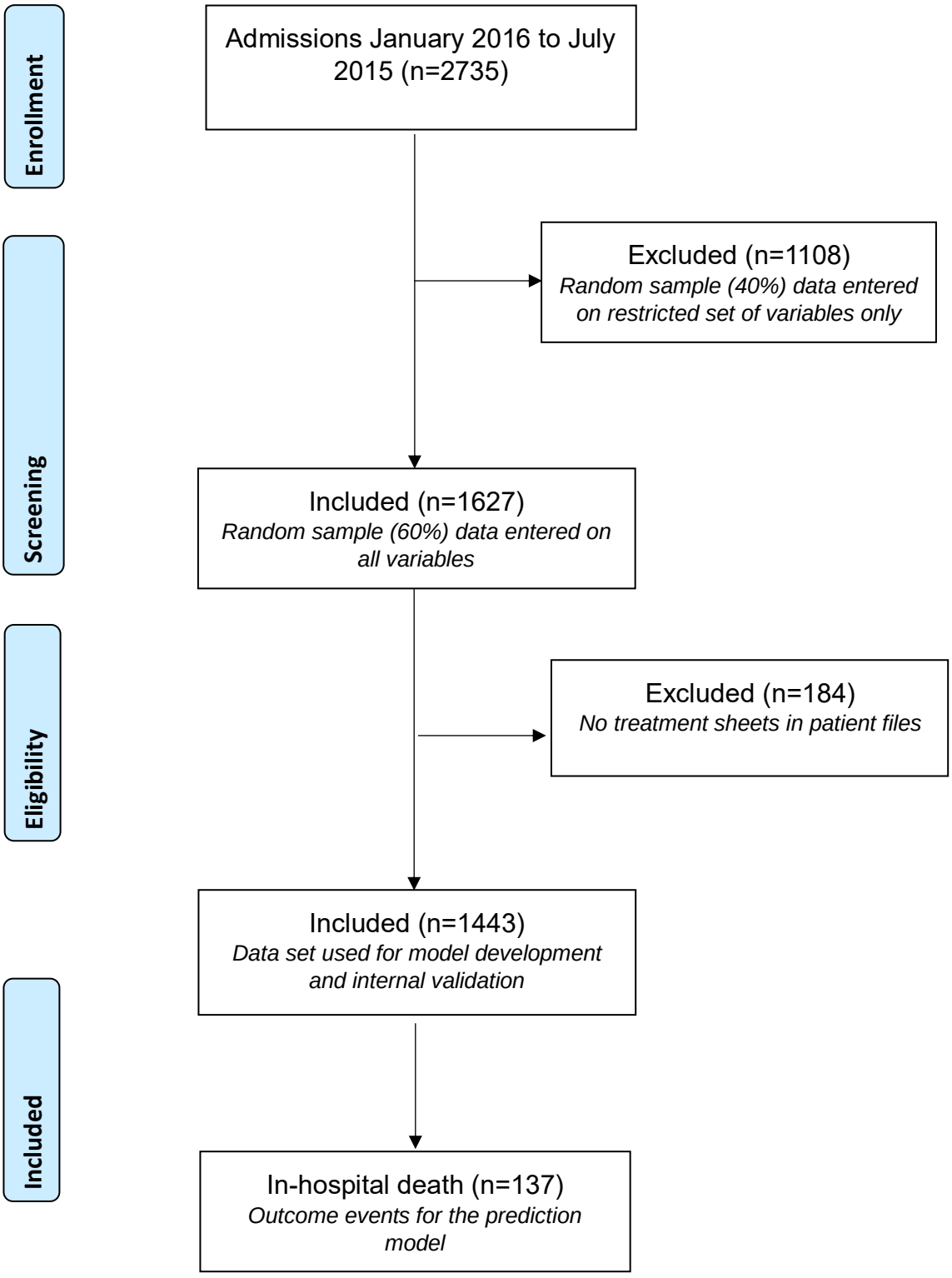



***Figure 5.2**: Patient flow chart for NETS model external validation*

*c) Outcome and Predictors*

The definition of outcome (in-hospital mortality) and predictors (birthweight, sex and five essential treatments) was the same as for model derivation (Section 5.2.1 (c) and (d)).

*d) Sample size*

At least 100 events and 100 non-events are recommended for external validation.(136) In the external validation dataset there were 137 events (in-hospital death) and 1300 non-events (left the neonatal unit alive).

*e) Missing data*

Considerations for missing data were the same as for model derivation described in Section 5.2.1(f). Multiple imputation of missing values was not conducted but descriptive analyses comparing missingness in the external validation data set against the derivation data set are summarised in Table 5.11 in the results section.

*f) Statistical analysis methods*

I applied the model coefficients obtained at derivation to the external validation data set to compute the predicted probability of in-hospital death for each patient. This was done using the built-in R function “predict.lm”. I thereafter estimated model performance at external validation in terms of discrimination and calibration by constructing a calibration curve as described in section 2.6.5 of Chapter 2. The final step was updating the original model to improve performance in the external validation data set.

Updating was done by adjustment of the NETS model intercept. This adjustment makes the mean predicted probability equal to the observed outcome frequency. To do this, I computed a correction factor which I then added to the original intercept to give a new intercept. The effect of this on model performance is to correct the calibration intercept (the calibration slope and predictor coefficients of the original model are retained).(136, 275) I finally evaluated the performance of the updated model by once again constructing a calibration curve. The code for external validation is included in Appendix 9.

## 5.3 Results

### *5.3.1 Clinical characteristics of the study patients*

***Table 5.4**: Clinical characteristics of neonates included in the model development data set*

| Characteristic | Random sample(n=5427) | | NETS* development data set(n=4840) | | Excluded from NETS (n=587)† | |
|---|---|---|---|---|---|---|
| | n | % | n | % | n | % |
| **1.Sex** | | | | | | |
| Male | 2937 | **54** | 2605 | **54** | 332 | **57** |
| Missing | 13 | **0.2** | 12 | **0.3** | 1 | **0.2** |
| **2.Birthweight (kg)** | | | | | | |
| <1 | 32 | **0.6** | 31 | **0.6** | 1 | **0.2** |
| 1-<1.5 | 136 | **3** | 115 | **2** | 21 | **4** |
| 1.5-<2.5 | 1180 | **22** | 1043 | **22** | 137 | **23** |
| 2.5-4 | 3841 | **71** | 3438 | **71** | 403 | **69** |
| >4 | 229 | **4** | 204 | **4** | 2 | **4** |
| Missing | 9 | **0.2** | 9 | **0.2** | 0 | |
| **3.Mode of delivery** | | | | | | |
| Spontaneous vaginal | 3011 | **57** | 2697 | **57** | 314 | **55** |
| Assisted vaginal | 3 | **0.1** | 1 | **0.02** | 2 | **0.3** |
| Breech | 43 | **1** | 40 | **1** | 3 | **1** |
| Caesarean section | 2212 | **42** | 1999 | **42** | 255 | **44** |
| Missing | 158 | **3** | 145 | **3** | 13 | **2** |
| **4.Outborn‡** | | | | | | |
| Yes | 123 | **2** | 107 | **2** | 16 | **3** |
| Missing | 0 | | 0 | **-** | 0 | **-** |
| **5. Apgar score (5min)** | | | | | | |
| 0-3 | 116 | **3** | 112 | **2** | - | **-** |
| 4-6 | 602 | **13** | 593 | **13** | - | **-** |
| 7-10 | 3992 | **85** | 3918 | **85** | - | **-** |
| Missing | 717 | **11** | 217 | **4** | 500 | **86** |
| **7.HIV exposure** | | | | | | |
| Exposed | 319 | **6** | 287 | **6** | 32 | **6** |
| Missing | 305 | **6** | 277 | **6** | 28 | **5** |
| **8.Outcome** | | | | | | |
| Discharged alive | 4838 | **89** | 4318 | **89** | 520 | **89** |
| Dead | 508 | **9** | 447 | **9** | 61 | **10** |
| Referred | 62 | **1** | 56 | **1** | 6 | **1** |
| Missing | 19 | **0.4** | 19 | **0.4** | 0 | **-** |

***** NETS: Neonatal Essential Treatment Score
† Excluded due to missing treatment sheets in the patient file
‡ Outborn refers to neonates admitted to the unit having been born either in another facility, at home or on the way to hospital

Table 5.4 shows the distribution of clinical characteristics in all the patients eligible for inclusion in model development(n=5427), those included (n=4840) and those

excluded (n=587). As shown in Figure 5.1, the reason for exclusion was absence of a treatment sheet in the patient file. The clinical characteristics are similar in the three groups shown in Table 5.4. This suggests that the treatment sheets were missing completely at random (MCAR) and therefore the model development data set (n=4840, 445 deaths) is representative of the population of eligible patients (n=5427, 508 deaths).

*5.3.2 Model derivation*

*a) Distribution of predictors in the derivation data set*

The distribution of the treatments included as predictors in the NETS model by in-hospital mortality (the outcome) are shown in Table 5.5. These treatments were defined and described in Table 5.2

***Table 5.5**: Distribution of predictors by in-hospital mortality in the NETS derivation dataset*

| Predictor*(n=4840) | Prescribed | | In-hospital death† | | Missing outcome | |
|---|---|---|---|---|---|---|
| | n | % | n | % | n | % |
| **1.Intravenous antibiotics** | 1892 | **39** | 383 | **20** | 4 | **0.2** |
| **2.Intravenous fluids** | 869 | **18** | 295 | **34** | 1 | **0.1** |
| **3.Enteral feeds** | 1381 | **29** | 11 | **1** | 8 | **0.6** |
| **4.Oxygen** | 295 | **6** | 120 | **41** | 0 | **0** |
| **5.Phenobarbital** | 69 | **1** | 42 | **61** | 0 | **0** |
| **6.None prescribed** | 834 | **17** | 18 | **2** | 8 | **1** |

***** Prescribed at admission

† Distribution of in-hospital deaths amongst patients who had prescriptions of treatments included as predictors in the NETS model. Denominator excludes those missing outcome

Individual patients had a varying combination of prescriptions of these treatments. The highest number of prescriptions was for intravenous antibiotics 1892/4840 (39%) but the highest proportion of deaths was for those who had a prescription of phenobarbital 42/69 (61%). However, 834/4840 (17%) neonates did not have a prescription of any of these treatments. The last column of the table (missing

outcome) shows the number and proportion for whom the outcome at the point of exit from the unit (death, referral or discharge alive) was not documented.

*b) Model specification*

The logistic regression model (lrm) specified to predict in-hospital mortality is shown in Box 5.1 and the detailed code is included in Appendix 9.

```
lrm (formula = outcome ~ birthweight + sex + antibiotics + fluids + feeds +
               oxygen + phenobarbital, data = moddata, x = T,
               y = T,linear.predictors = T)
```

***Box 5.1**: Logistic regression model for the NETS score*

A complete case regression analysis approach as described in Section 5.2.1 (f) was used. This means that all the observations with a missing variable were dropped. A total of 19 were missing outcome, 9 missing birthweight and 12 gender for a total of 40 observations. The derivation model thus included 4800 observations with 445 deaths (EPV=44).

*c) Model Estimation*

The coefficients estimated for each predictor from the multivariable logistic regression model in the log odds scale are shown in Table 5.6.

***Table 5.6**: Coefficients from the logistic regression model for the NETS score*

| Model parameter | Coefficients * | Standard Error | P value |
|---|---|---|---|
| **Intercept** | -4.152 | 0.178 | <0.0001 |
| **ELBW(<1kg)** | 5.684 | 1.051 | <0.0001 |
| **VLBW(1-<1.5kg)** | 4.536 | 0.313 | <0.0001 |
| **LBW(1.5-<2.5kg)** | 1.419 | 0.144 | <0.0001 |
| **Macrosomia(>4kg)** | -0.293 | 0.516 | 0.5706 |
| **Male** | -0.313 | 0.133 | 0.0184 |
| **Intravenous antibiotics** | 1.370 | 0.186 | <0.0001 |
| **Intravenous fluids** | 1.326 | 0.158 | <0.0001 |
| **Enteral feeds** | -1.914 | 0.332 | <0.0001 |
| **Oxygen** | 0.614 | 0.176 | 0.0005 |
| **Phenobarbital** | 2.595 | 0.306 | <0.0001 |

* Values in log odds scale for the predictors with in-hospital mortality as the outcome
ELBW: Extremely Low Birth Weight, VLBW: Very Low Birth Weight, LBW: Low Birth Weight

The intercept is the baseline log odds of in-hospital mortality for patients who did not have a prescription for any of these treatments and belonged to the reference category for birth weight (normal weight, 2.5kg to 4kg) and sex (female). These coefficients were summated to give the linear predictor of in-hospital death as shown in Box 5.2.

```
Linear predictor = -4.8711 + 5.6836*ELBW + 4.5359*VLBW + 1.4186*LBW –
                   0.2927*Macrosomia – 0.3125*Male + 1.3695*Antibiotics +
                   1.3256*Fluids – 1.9135*Feeds + 0.6142*Oxygen +
                   2.5947*Phenorbabital
```

***Box 5.2**: NETS derivation regression formula for the linear predictor of in-hospital mortality*

The linear predictor was then converted to predicted probability of in-hospital mortality using the formula described in Box 2.3 of Chapter 2 and the model performance evaluated. The results are presented in the next section.

*d) Model apparent performance*

As defined in Chapter 2, the apparent performance of the model refers to the predictive accuracy of the model in the derivation dataset. In this case it is the predictive accuracy of the NETS model in predicting in-hospital mortality for the 4800 observations (445 deaths) in the derivation dataset. I first present the measures of overall model performance ($R^2$ and Brier score) followed by estimates of discrimination and calibration.

*Overall performance*

Table 5.7 displays the overall measures of the NETS model.

***Table 5.7**: Overall measures of NETS model apparent performance*

| Index | Result | Scale | Interpretation |
|---|---|---|---|
| **Nagelkerke $R^2$** | 0.511 | 0-1 | Higher when those without outcome are closer to 0% predicted probability and those with outcome closer to 100% |
| **Brier score** | 0.050 | 0-1 | 0 perfect discrimination |

The Nagelkerke $R^2$ ranges from 0-1, the greater the value the better as the closer the predicted probabilities are to the observed.(136) However $R^2$ values for logistic regression are usually low.(276) In contrast the smaller the value of the Brier score the better the predictive performance with a score of zero representing perfect prediction. (136) However, these two overall measures do not distinguish between

the discrimination and calibration aspects of model performance.(277) In the next section, I present model discrimination and calibration using separate measures.

*Discrimination and calibration*

The calibration plot in Figure 5.3 shows estimates for NETS model discrimination and calibration in the derivation data set.

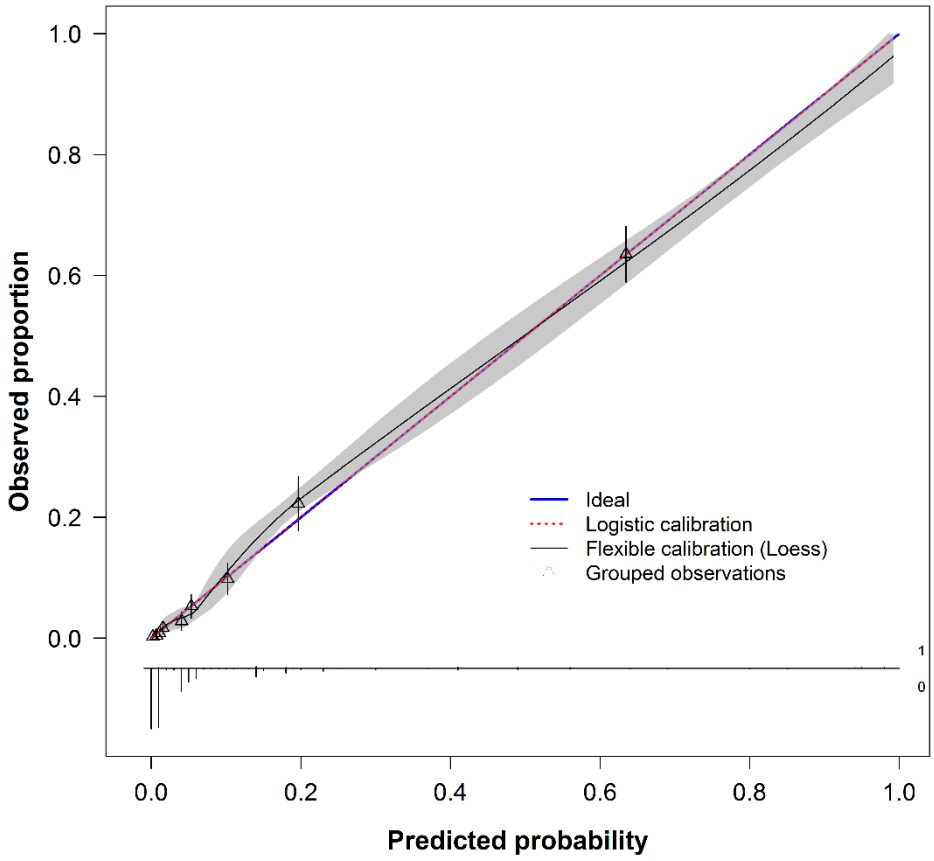



***Figure 5.3**: Calibration curve for NETS in the derivation data set*

The NETS logistic regression model computes the probability of in-hospital death for each patient in the data set from their combination of model predictors. This is done by summing the product of the predictor values for each patient by the corresponding regression coefficient using the formula shown in Box 5.2. The

result for each patient is a predicted probability of in-hospital death (on a scale from 0 to 1).

In Figure 5.3, the predicted probability of in-hospital death is plotted on the X-axis against the observed proportion on the Y-axis. Since the observed outcome is binary (dead or alive), the observed proportion was obtained by grouping the observations into tenths of predicted probability of in-hospital death. These tenths are represented by the triangles plotted on Figure 5.3 (grouped observations). Within each tenth the proportion of patients who died in hospital was then computed on a 0 to 1 scale and the result plotted on the Y-axis.

Discrimination is assessed by the c-statistic of the logistic regression model. This is equivalent to the area under the receiver operating curve(AUROC). The c-statistic for NETS in the derivation data set was 0.92 (95% C.I 0.90 to 0.93). This means that using the NETS model, the predicted probability of in-hospital death for patients who (observed death) was higher than the predicted probability in those who left the unit alive 92% of the time. This quantifies the ability of the NETS model to tell apart patients who died in hospital and those who left alive by assigning higher predicted probabilities of in-hospital death to the former.

Calibration is quantified by two measures; the intercept and the slope. For the NETS model at derivation, the intercept was 0 and slope 1. Regression models always give an intercept of zero and slope of one (perfect calibration) for the derivation data set and this is shown in Figure 5.2.(135) These parameters measure the extent of agreement between the NETS predicted probability of in-

hospital mortality and the observed in-hospital mortality. The calibration intercept compares the mean predicted in-hospital death to the observed incidence. The intercept will be greater than zero if mean predicted mortality is less than the observed proportion and less than zero if it is greater than the observed. The slope in Figure 5.3 shows how well the predicted probabilities match the observed across the tenths using two different methods: logistic calibration and Loess smoothing. The interpretation of both is however the same i.e. for perfect calibration the slope is equal to one.

After deriving the NETS model, I undertook an internal validation to examine and quantify any optimism. I will now describe the results of this assessment in the next section.

*5.3.3 Internal validation*

*a) Correction of model performance for optimism*

Table 5.8 shows calibration of the NETS model after correction for optimism (Chapter 2, Section 2.6.4). The original index refers to the model's apparent performance in the derivation data set as was shown in Table 5.7. The training and test columns represent performance in the bootstrap samples as described in Chapter 2 (Box 2.4). Optimism for each parameter is computed by subtracting the test performance from the training performance. The original performance indices are then corrected by subtracting the optimism for each respective parameter to yield the corrected indices, termed 'Corrected index' in Table 5.8.

***Table 5.8**: Calibration of the NETS model after correction for optimism*

| | Original index | Training | Test | Optimism | Corrected index |
|---|---|---|---|---|---|
| **Intercept*** | 0.000 | 0.000 | -0.062 | 0.062 | -0.062 |
| **Slope*** | 1.000 | 1.000 | 0.979 | 0.022 | 0.979 |

* Calibration intercept and slope

There is minimal change in the indices after correction because the values of optimism are small. This reflects the large sample size, small number of predictors and no variable selection (445 in-hospital deaths, EPV=44) used for model derivation which meant there was little overfitting during model estimation. The same process was applied to computing the optimism corrected c-statistic as shown in Table 5.9.

***Table 5.9**: Discrimination of the NETS model after correction for optimism*

| Original | Training | Test | Optimism | Corrected |
|---|---|---|---|---|
| **0.918** | 0.918 | 0.916 | 0.002 | 0.916 |

Like the overall performance measures in Table 5.8, there was a very slight change in the c-statistic after correction i.e. from 0.918 to 0.916.

*b) Updating the model coefficients after internal validation*

The presence of significant optimism indicates model overfitting which would require adjustment of the model coefficients to reduce deterioration in performance when the model is applied to an external population. From Tables 5.8 and 5.9 the NETS models displayed little optimism but I still went ahead to examine the extent to which the model coefficients would change if adjusted. This is referred to as coefficient shrinkage after estimation.(136)

The original coefficients are adjusted by multiplying by a factor referred to as the "shrinkage factor". The shrinkage factor is given by the optimism corrected calibration slope shown in Table 5.8 (0.979). This figure is close to 1 and there is minimal change in the NETS model coefficients after multiplication as shown in Table 5.10.

***Table 5.10**: Shrinkage of NETS model coefficients to correct for optimism*

| Model parameter | Original* | Shrunk* |
|---|---|---|
| **Intercept** | -4.152 | -4.063 |
| **ELBW(<1kg)** | 5.684 | 5.562 |
| **VLBW(1-<1.5kg)** | 4.536 | 4.439 |
| **LBW(1.5-<2.5kg)** | 1.419 | 1.388 |
| **Macrosomia(>4kg)** | -0.293 | -0.286 |
| **Male** | -0.312 | -0.306 |
| **Intravenous antibiotics** | 1.370 | 1.340 |
| **Intravenous fluids** | 1.326 | 1.297 |
| **Enteral feeds** | -1.913 | -1.872 |
| **Oxygen** | 0.614 | 0.601 |
| **Phenobarbital** | 2.595 | 2.539 |

* Values on the log odds scale for the predictors with in-hospital mortality as the outcome

Given the small change in the model coefficients after shrinkage shown in Table 5.10, the original coefficients were taken forward for external validation. I now describe these results next.

### *5.3.4. External validation*

a) Participants

As shown in Figure 5.2, the external validation data set consisted of 1443 observations from the same facility as the derivation data set but from a more recent period, January 2016 to July 2016 (temporal external validation). Table 5.11 shows comparison of clinical characteristics of the patients included in the derivation data set with those in the external validation data set.

***Table 5.11**: Characteristics of patients included in NETS derivation and external validation*

| Characteristic | NETS* derivation (n=4840) | | NETS* external validation (n=1443) | |
|---|---|---|---|---|
| | n† | % | n† | % |
| **1.Sex** | | | | |
| Male | 2605 | **54** | 850 | **59** |
| Missing | 12 | **0.3** | 2 | **0.1** |
| **2.Birthweight (kg)** | | | | |
| <1 | 31 | **0.6** | 10 | **1** |
| 1-<1.5 | 115 | **2** | 40 | **3** |
| 1.5-<2.5 | 1043 | **22** | 316 | **22** |
| 2.5-4 | 3438 | **71** | 1002 | **69** |
| >4 | 204 | **4** | 74 | **5** |
| Missing | 9 | **0.2** | 1 | **0.1** |
| **3.Mode of delivery** | | | | |
| Spontaneous vaginal | 2697 | **57** | 897 | **63** |
| Assisted vaginal | 1 | **0.02** | 0 | **0** |
| Breech | 40 | **1** | 19 | **1** |
| Caesarean section | 1999 | **42** | 509 | **36** |
| Missing | 145 | **3** | 18 | **1** |
| **4.Outborn ‡** | | | | |
| Yes | 107 | **2** | 57 | **4** |
| Missing | 0 | **0** | 0 | **0** |
| **5. Apgar score (5)** | | | | |
| 0-3 | 112 | **2** | 28 | **2** |
| 4-6 | 593 | **13** | 192 | **14** |
| 7-10 | 3918 | **85** | 1122 | **84** |
| Missing | 217 | **4** | 69 | **5** |
| **7.HIV exposure** | | | | |
| Exposed | 287 | **6** | 74 | **5** |
| Missing | 277 | **6** | 80 | **6** |
| **8.Outcome** | | | | |
| Discharged alive | 4318 | **89** | 1262 | **88** |
| Dead | 447 | **9** | 137 | **10** |
| Referred | 56 | **1** | 38 | **3** |
| Missing | 19 | **0.4** | 6 | **0.4** |
| **9. Admission Diagnosis** | | | | |
| Birth Asphyxia | 2011 | **42** | 494 | **34** |
| Preterm/LBW | 979 | **201** | 257 | **18** |
| Neonatal sepsis | 495 | **10** | 172 | **12** |
| Respiratory D. Syndrome | 331 | **7** | 103 | **7** |
| Neonatal jaundice | 272 | **6** | 141 | **10** |
| Others | 513 | **11** | 332 | **23** |

* NETS: Neonatal Essential Treatment Score

† Denominators for the variables obtained by subtracting the missing data from the sample (4840 for derivation, 1443 for external validation)

‡ Outborn refers to neonates admitted to the unit having been born either in another facility, at home or on the way to hospital

Possibly notable differences in the distribution of clinical characteristics between the external validation data set in comparison to the derivation data set were: males (59% vs 54), caesarean section (36% vs 42%), outborn (4% vs 2%) and admission diagnosis other than the top 5 conditions (23% vs 11%).

Table 5.12 shows the distribution of the essential treatments used as predictors by in-hospital mortality in the external validation data set and the derivation data set. This table explores variation in clinical practice in the two different time periods (derivation vs external validation) with respect to prescription of the essential treatments selected as predictors for the NETS model.

***Table 5.12**: Distribution of predictors by in-hospital mortality*

| **Predictors** | **NETS* External validation(n=1443)** | | | | | | **NETS* Development (n=4840)** | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **Prescribed** | | **Dead†** | | **Missing§** | | **Prescribed** | | **Dead†** | | **Missing§** | |
| | **n** | **%** | **n** | **%** | **n** | **%** | **n** | **%** | **n** | **%** | **n** | **%** |
| **Intravenous antibiotics** | 780 | **54** | 123 | **16** | 1 | **0.1** | 1892 | **39** | 383 | **20** | 4 | **0.2** |
| **Intravenous fluids** | 495 | **34** | 116 | **23** | 1 | **0.2** | 869 | **18** | 295 | **34** | 1 | **0.1** |
| **Enteral feeds** | 598 | **41** | 7 | **1** | 2 | **0.3** | 1381 | **29** | 11 | **1** | 8 | **0.6** |
| **Oxygen** | 105 | **7** | 105 | **24** | 0 | **0** | 295 | **6** | 120 | **41** | 0 | **0** |
| **Phenobarbital** | 72 | **5** | 21 | **25** | 1 | **1.5** | 69 | **1** | 42 | **61** | 0 | **0** |
| **None** | 196 | **14** | 9 | **5** | 3 | **1.5** | 834 | **17** | 18 | **2** | 8 | **1** |

* NETS: Neonatal Essential Treatment Score
† Denominators obtained by subtracting the number missing outcome data for each predictor
§ Missing outcome data for each predictor

There was a higher proportion of prescription of all the five essential treatments in the external validation data set compared to the derivation data set. A smaller proportion of deaths was observed in those who had a prescription of intravenous antibiotics, intravenous fluids, oxygen and phenobarbital in the external validation data set. This may be a result of perhaps a lowering of the severity of illness

threshold for prescribing interventions. The proportion of patients missing documentation of outcome was however similar.

c) Model performance at external validation

Figure 5.4 shows NETS model performance in the external validation data set.

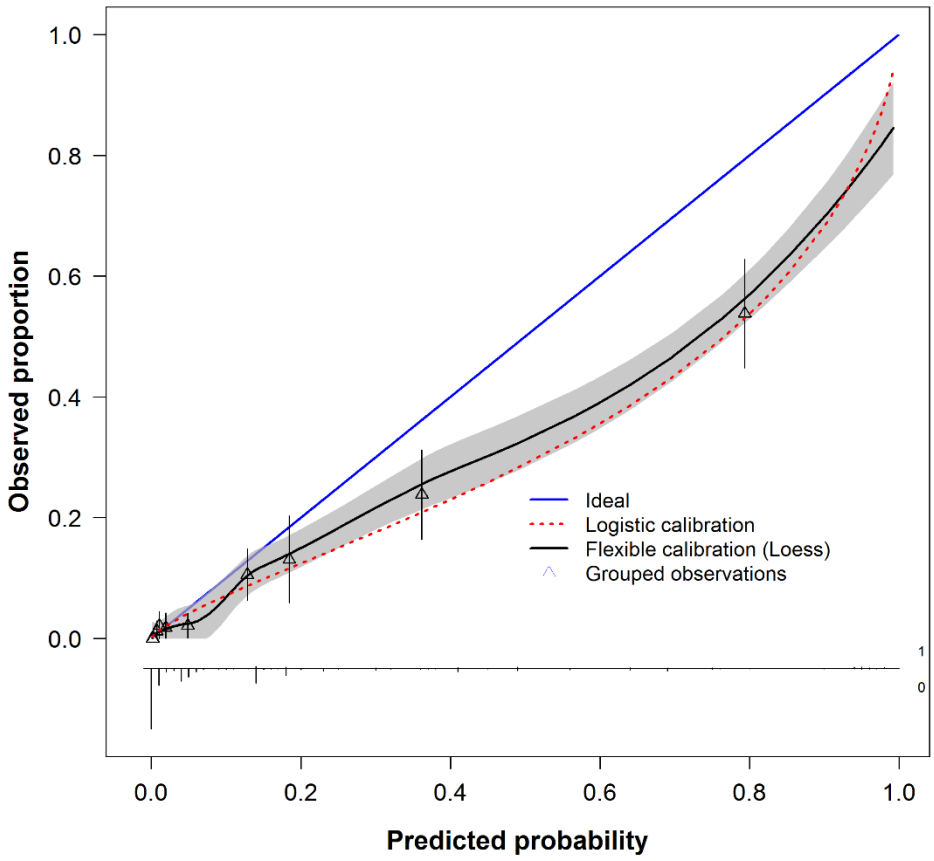



***Figure 5.4**: Calibration curve for NETS in the external validation data set*

The model performance has deteriorated relative to that observed in the derivation data set as shown in Figure 5.2. Discrimination has slightly dropped from a c-statistic of 0.92 to 0.89 (95% C.I 0.86 to 0.92). A more pronounced change is seen in the calibration. The calibration intercept has declined from 0 to -0.72 (95% C.I -0.96 to -0.49). Since it is less than the ideal zero, it means the mean predicted in-

hospital death is greater than the mean observed and therefore the model over predicts in-hospital mortality in the new data set. With regards to the calibration slope, the new slope is 0.76 (95% C.I 0.65 to 0.87) down from 1. The model curves run below the ideal line (perfect calibration) again demonstrating that the model is over-predicting across the deciles of observed in-hospital death. This is more pronounced at higher predicted probabilities of in-hospital death.

d) Model updating after external validation

Due to the observed deterioration in performance shown in Figure 5.4, I updated the NETS model to improve performance in the new data set. I did this by adjusting the original model intercept as described in Section 5.2.3 (f). The updated intercept coefficient is -4.877 while the original was -4.152 (see Table 5.6). The predictor coefficients were not altered. This final model is shown in Table 5.13.

***Table 5.13**: Final NETS model after updating*

| Model parameters | Coefficients* | P value |
|---|---|---|
| **Intercept** | -4.877 | <0.0001 |
| **ELBW(<1kg)** | 5.684 | <0.0001 |
| **VLBW(1-<1.5kg)** | 4.536 | <0.0001 |
| **LBW(1.5-<2.5kg)** | 1.419 | <0.0001 |
| **Macrosomia(>4kg)** | -0.293 | 0.5706 |
| **Male** | -0.312 | 0.0184 |
| **Intravenous antibiotics** | 1.370 | <0.0001 |
| **Intravenous fluids** | 1.326 | <0.0001 |
| **Enteral feeds** | -1.913 | <0.0001 |
| **Oxygen** | 0.614 | 0.0005 |
| **Phenobarbital** | 2.595 | <0.0001 |

* Values in log odds scale for the predictors with in-hospital mortality as the outcome

e) Performance of final model

The calibration curve constructed to evaluate performance after updating the NETS model coefficient is shown in Figure 5.5.

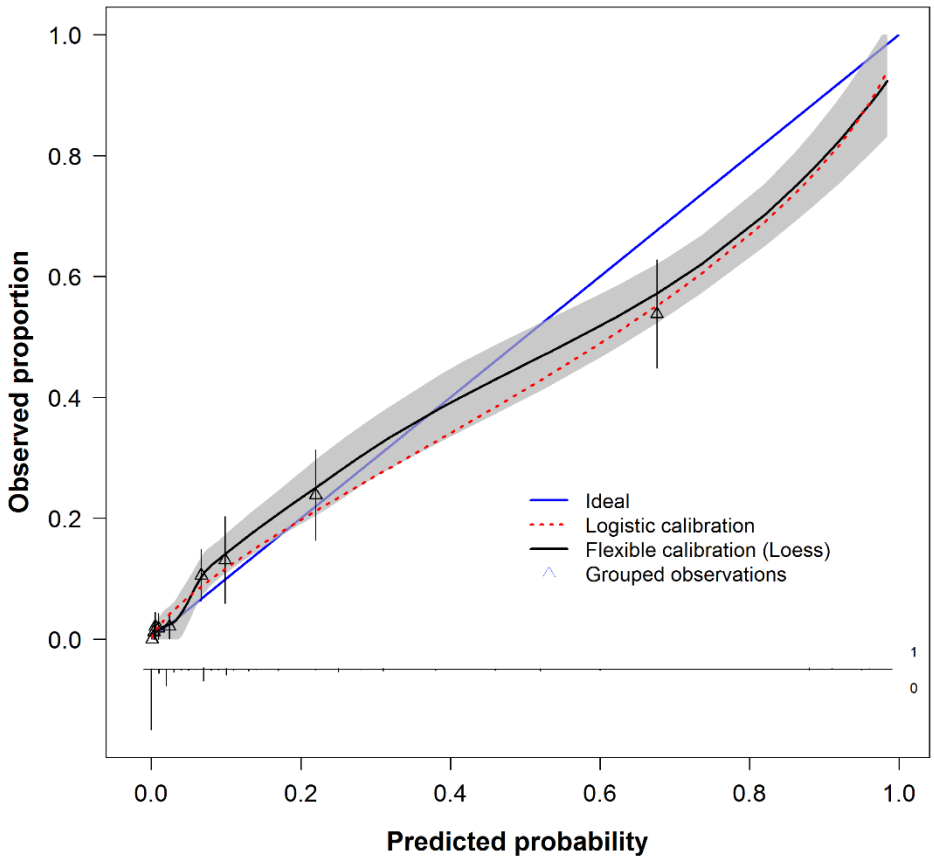



***Figure 5.5**: Calibration curve for NETS model after external validation and updating*

In Figure 5.5 mean predicted in-hospital mortality now matches the observed with calibration intercept 0 (95% C.I -0.24 to 0.24). The calibration slope however remains 0.76(95% C.I 0.65 to 0.87) as only the intercept was adjusted. By adjusting the intercept, the whole of the calibration curve is shifted upwards and therefore predicted probabilities across all the tenths are now closer to the observed as represented by the ideal line. Discrimination measured by the c-statistic remains the same at 0.89.

## 5.4 Discussion

### *5.4.1 Summary of findings*

In this section I limit discussion to the specific findings for this Chapter. Further discussion of these results considering all the thesis results is presented in Chapter 8. In this Chapter, I have developed the Neonatal Essential Treatment Score (NETS) that is based on 5 essential treatments prescribed at the time of admission: intravenous antibiotics, intravenous fluid infusion, enteral feeding, supplemental oxygen and phenobarbital. In addition, I included birthweight and sex. Whereas birthweight is a continuous variable, it was included in the model as an ordinal categorical variable based on the 5 established clinical categories. All the other predictors were binary. There was no predictor selection procedure during model estimation as these predictors were based on clinical criteria.

The predictive performance of NETS at derivation, internal validation and external validation is summarised in Table 5.14. In external validation, the final c-statistic is 0.89 while calibration slope is 0.76.

*Table 5.14: Summary of the predictive accuracy of NETS*

| **Parameter** | **Metric** | **Results by modelling stage** | | |
|---|---|---|---|---|
| | | Derivation | Internal validation | External validation |
| **1. Discrimination** | c-statistic | 0.917 | 0.916* | 0.89† |
| **2. Calibration** | intercept | 0 | -0.063* | 0† |
| | slope | 1 | -0.979* | 0.76† |

* Optimism corrected

† Final model after updating

*5.4.2 Model specification and estimation in derivation*

The treatments included as predictors in the NETS model were selected based on clinical criteria to reflect in-patient essential newborn care in low resource setting hospitals. This is similar to the approach used to select the neonatal intensive care treatments included in the two model derivation studies included in the systematic review presented in Chapter 3.(104, 191) As was the case for NETS , this method of predictor selection coupled with the full model approach is particularly useful in instances where the list of potential predictors is short and they are collected routinely.(278)

A key difference between NETS and these earlier models however, is the handling of the treatments during model estimation. Gray et al and Shah et al as discussed in Chapter 3 used expert panels to assign weights to the treatments which were then summated to give a score for each patient.(104, 191) This score was then used as a predictor. In NETS, the weights assigned to the predictors were the coefficients obtained from the logistic regression model. This approach is transparent and reproducible. In addition, in NETS the treatments were used as individual predictors rather than summating all the weights.

*5.4.3 Predictive performance at derivation*

The ability of the NETS model obtained at the derivation stage (original model) to distinguish between neonates who died in hospital and those who did not die is quantified by the c-statistic. The c-statistic was 0.92 (95% C.I 0.90-0.93). This represents good discriminatory ability as the point estimate (0.92) and the lower bound of the 95% confidence interval are both above the threshold of 0.80 where

discrimination is considered adequate.(153) The calibration intercept and slope were 0 and 1. By definition, regression modelling always results in an intercept of 0 and slope of 1 at model derivation thus limiting the usefulness of this metric at this stage.(135, 277)

Three of the studies included in Chapter 3 reported model discrimination for in-hospital mortality using area under the receiver operating curve (AUROC). The NETS c-statistic at derivation, 0.92(95% C.I 0.90 to 0.93), is higher than 0.82±0.04 reported by Eriksson et al.(193) The articles by Oygur et al and Wu et al only included neonates weighing less than 1500g thus barring direct comparison of predictive performance against NETS where I included all birthweight categories.(197, 198) None of the studies reviewed in Chapter 3 reported model calibration intercept or slope.

*5.4.4 Predictive performance at internal validation*

The aim of internal validation was to estimate the extent to which the NETS model fits the peculiarities in the distribution of the predictors as observed in the derivation data set that may not be observed in an external data set. This is referred to as model overfitting as described in Chapter 2 (Section 2.6.4) and is more likely to occur where the sample size is small with many predictors relative to the observed outcome of interest (events per variable, EPV).

I used the bootstrap approach which is the recommended method for estimating the extent of optimism in the model's apparent performance at derivation stage.(136, 272, 279) The optimism adjusted c-statistic was 0.916 compared to the

original 0.917. Similarly, there was minimal change in the NETS model coefficients after shrinkage to adjust for optimism. Shrinkage refers to reduction of the magnitude of the coefficients by multiplying by a shrinkage factor. The factor is the calibration slope of the model after internal validation and ranges from 0 to 1. Given that the shrunk coefficient is a product of the shrinkage factor and original coefficient, the large shrinkage factor obtained for the NETS model (0.97) explains the minimal change observed. There were 40 in-hospital deaths per NETS model predictor (against the recommended 10 to 20) and this large sample size explains the minimal overfitting demonstrated by internal validation. I therefore retained the original model for external validation.

No comparisons could be made between the results of internal validation I report here and the 10 studied included in the systematic review in Chapter 3 as none performed internal validation. As I discussed in Chapter 3, this is a significant limitation of existing literature on neonatal treatments scores.

*5.4.5 Performance at external validation*

For external validation, I examined the temporal transportability of NETS in the study hospital. There was a greater reduction in both the NETS discrimination and calibration in comparison to the changes observed at internal validation. The c-statistic declined to 0.89 from 0.92 but there was overlap in the confidence intervals, 0.86 to 0.92 vs 0.90 to 0.93 for external validation and original models respectively. The calibration intercept reduced from 0 (95% C.I -0.13 to 0.13) to -0.72 (-0.96 to -0.49) demonstrating that mean predicted in-hospital death was greater than the mean observed in the external validation data set. This same

applies to calibration across the tenths of predicted risk with the calibration slope declining from 1(95% C.I 0.92 to 1.08) to 0.76 (0.65 to 0.87).

The deterioration in predictive accuracy at external validation may be explained by differences between the derivation and the validation populations. Three types of differences have been described: (1) differences in definition and measurement of predictors and outcomes, (2) differences in the patients included (case-mix) and (3) inclusion of fewer patients in validation than derivation.(280)

With regards to predictor and outcome definition, the same data collection procedures were used for the derivation and validation set therefore this factor was less likely to be applicable. There were however differences in the study populations as shown in Table 5.12. This table shows differences in the distribution of the essential treatments (NETS model predictors) overall and their distribution by in-hospital mortality. This may represent a real difference in case-mix for example due to change in admission practices for less severely ill neonates or change in prescribing practices. It may also be related to the third category of factors explaining deterioration in model performance i.e. difference in population size. The validation data set was smaller (1443 vs 4840) and spanned a shorter period (7 months vs 18 months). It is therefore conceivable that the case-mix difference observed is random variation. Although secular change in clinical practice cannot be ignored as hospital personnel do change quite often in Kenya e.g. following interdepartmental transfers and arrival of new staff.

None of the studies included in the systematic review in Chapter 3 conducted external validation thus precluding comparison with NETS.

### *5.4.6 Model updating*

To remedy the deterioration in model performance, I explored updating the model by adjusting the intercept. This resulted in the final model with c-statistic 0.89 (95% C.I 0.86 to 0.92), calibration intercept 0 (95% C.I -0.24 to 0.24) and calibration slope 0.76 (95% C.I 0.65 to 0.87). Model updating allows the incorporation of new information from the external validation data set to the original model. This is likely to improve the generalizability of the model and is preferable to the development of multiple models to predict the same outcome in different populations.(280)

### *5.4.7 Final model and clinical utility*

The final model after external validation expressed as a formula is shown in Figure 5.5 based on the generic formula shown in Box 2.2 of Chapter 2. The coefficients are derived from the final NETS model shown in Figure 5.13

```
Linear predictor = -4.8711 + 5.6836*ELBW + 4.5359*VLBW + 1.4186*LBW –
                     0.2927*Macrosomia – 0.3125*Male + 1.3695*Antibiotics +
                     1.3256*Fluids – 1.9135*Feeds + 0.6142*Oxygen +
                     2.5947*Phenobarbital
```

***Box 5.3**: Final logistic regression model for NETS after updating the intercept*

Box 5.6 presents the formula used to estimate the predicted probability of in-hospital death for each individual patient. This is obtained by summating the predictor coefficients to give the odds of death in the log scale. Using the formula described in Box 2.3 of Chapter 2, this log odd is then converted to a probability.

The coefficient for the reference category for each predictor is zero. For birthweight, the reference is normal weight (2.5-4kg) and for sex it is female. The rest of the predictors are treatments and binary so the reference is that the treatment was not prescribed. The clinical application of this formula and its potential utility is further described in Chapter 7 using decision curve analysis.

### *5.4.8 Limitations*

Clinical practice changes with time and therefore treatments prescribed may change. However, for the case of NETS the intervention category represented by the predictor will likely remain constant. Taking the example of intravenous antibiotics, while the specific first line antibiotics may change from penicillin and gentamicin to other medicines it is very likely that there will always be recommended first line intravenous antibiotics. Nevertheless, all prediction models require updating over time and introduction of new interventions e.g. BCPAP can be incorporated in the updated models.(269)

If the expected reduction of in-hospital neonatal mortality is realised, death will become a rarer outcome and the NETS model may no longer be a valid predictor of death. However, as I demonstrated in Chapter 4 reduction of mortality is likely to be accompanied by increase in morbidity e.g. longer length of stay. In such a situation, the NETS model may be updated to predict a chosen morbidity or a composite outcome that includes in-hospital mortality.

I only carried out temporal external validation of the NETS model. This does not translate to its generalisability in other neonatal units offering essential neonatal

care as this can only be assessed by externally validating the model with such external data (geographical external validation). However, there exists within Kenya the potential to conduct geographical external validation using routinely collected data from other neonatal units.(281-283) This offers a line for future research on the NETS model.

### *5.4.9 Implications*

Essential treatments prescribed at admission obtained from routine inpatient data can be used to predict in-hospital mortality in a low resource setting neonatal unit. As treatment data is often available from routine records the NETS model can therefore feasibly be used in routine clinical practice in such settings. Generalisability of the NETS model will however require further assessment using data from similar neonatal units (geographical external validation).

# Chapter 6 : Development and validation of the Score of Essential Neonatal Symptoms and Signs (SENSS)

## 6.1 Introduction

### *6.1.1 Background*

In Chapter 5, I presented development and validation of the NETS model. In prognostic modelling research, head to head comparison of prognostic models in the same population is rarely done. Such comparisons allow the evaluation of the performance of different models to inform the selection of the most appropriate model for the population of interest.(148) An example of the utility of comparisons is illustrated by the evaluation of the Framingham risk score versus the QRISK model for predicting 10-year cardiovascular disease risk in the UK population.(284)

Collins and colleagues compared the performance of the UK National Institute of Health and Care Excellence (NICE) version of the Framingham risk score, the QRISK1 risk score and the QRISK2 in The Health Improvement Network (THIN) data set.(284) At the time the study was conducted, NICE recommended the use of the Framingham risk score despite evidence suggesting that QRISK performed better.(285) The primary outcome was 10-year risk of the first diagnosis of a cardiovascular event (myocardial infarction, coronary heart disease, stroke, and transient ischaemic attack). Model performance was measured using discrimination and calibration. The independent authors concluded that the QRISK2 risk score has a marginally better performance compared to QRISK1. However, they demonstrated that both QRISK2 and QRISK1 perform better than the NICE Framingham risk score. These results underscore the importance of an independent external evaluation of risk prediction models.

The ideal approach to model comparison is independent external validation of existing models.(284) However, for my DPhil work, none of the neonatal models I identified in Chapter 3 is fit for external validation in the data set I used for NETS. This is because the predictors needed to implement the models are largely unavailable in this data set and in most LMIC settings. However, my study data set includes signs and symptoms of severe illness which are used individually in routine clinical practice as predictors of severe illness as described in Chapter 1. These therefore provide an alternative approach to predicting in-hospital mortality.

In this Chapter, I therefore develop and validate a model to predict in-hospital mortality using signs and symptoms of severe illness as predictors; the Score for Essential Neonatal Symptoms and Signs (SENSS). I thereafter compare the performance of SENSS against that of NETS at external validation. This covers objective 3 of this thesis. In the next section, I will first review the evidence base for these signs and symptoms as clinical predictors.

### *6.1.2 Signs and symptoms of severe illness in neonates as prognostic factors*

Prognostic factor research is one of the key components of the Prognosis Research Strategy (PROGRESS) framework as I described in Chapter 1. Prognostic factor research aims to identify factors associated with a certain clinical outcome.(69) Regarding neonates and young infants (age 0-60 days) in low resource settings, Opiyo et al conducted a systematic review to identify the best clinical predictors of severe illness where diagnostic tests are not routinely available.(51) Severe illness was defined as cases requiring referral or in-hospital

care. This review identified five well conducted studies in resource limited settings which included 17,506 participants.

Eight symptoms and signs were identified as suitable clinical predictors and are summarised in Table 6.1. This table shows the range of odds ratios for each predictor for the outcome of severe illness identified in the five studies. Besides the statistical association, other criteria that were considered in their selection were prevalence in the enrolled infants in the primary studies and ease of clinical recognition.(51)

***Table 6.1**: Clinical predictors of severe illness in young infants (0-60 days)*

| Predictor | Number of studies | Odds Ratio (Range)† |
|---|---|---|
| **1. History of feeding difficulty*** | 3 | 1.5-10 |
| **2. History of convulsions*** | 2 | 1.5-15.4 |
| **3. Change in level of activity** | 1 | 1.5 |
| **4. Temperature ≥37.5°C** | 3 | 1.5-3.4 |
| **Temperature <35.5°C** | 1 | 9.2 |
| **5. Respiratory rate≥60 bpm** | 2 | 1.5-2.7 |
| **6. Severe chest indrawing*** | 4 | 1.5-8.9 |
| **7. Grunting** | 2 | 1.5-2.9 |
| **8. Cyanosis*** | 3 | 1.5-13.7 |

* Available in data set
† Point estimates of the odds ratio, range given where reported in more than one study

These clinical predictors are included in the standard admission record for neonates (Newborn Admission Record, NAR) which was the principal source document for the study data set (Appendix 10). Besides these eight predictors, the NAR also includes floppiness or inability to suck and bulging fontanelle. Bulging fontanelle does not improve the sensitivity and specificity of the eight in Table 6.1 (51) but is an independent predictor of bacterial meningitis in young infants (likelihood ratio 8.5).(286) Floppiness is considered a sign of serious systemic or neurological illness in clinical practice and together with inability to suck are

included in the NAR as measures of the level of consciousness.(287) I therefore considered these ten symptoms and signs as candidate prognostic variables.

Whereas the evidence as shown in Table 6.1 is for individual predictors of severe illness, I was interested in examining the utility of these clinical predictors in a prognostic model with in-hospital mortality as the outcome. The final steps I took to arrive at the final list of predictors are described in the methods section (Section 6.2).

### *6.1.3 Rationale*

The candidate clinical predictors identified are currently used for clinical decision making where the presence of any one indicates the need for hospital care which typically includes intravenous antibiotics, intravenous fluids, oxygen and enteral feeds.(51) However this approach as described in Chapter 1 (Table 1.5) prioritises sensitivity over specificity. (51) Nonetheless, from a prognostic modelling perspective there is robust evidence to support their use as clinical predictors of the need for hospital care and it is therefore reasonable to consider them as candidate prognostic variables for mortality. In addition, these predictors are included in the study data set signalling their potential availability from routine data obtained from similar settings.

However, available data shows that these signs and symptoms are often not completely documented.(32) Efforts are currently underway in Kenya to improve documentation of paediatric routine data in similar settings in Kenya.(281, 283, 288) With respect to this study, multiple imputation provides a suitable and efficient

approach to deal with missingness in the interim in order to answer the question; if these signs and symptoms are well documented, how would the performance of a model based on these predictors compare against the NETS model when evaluated in the same population?

### *6.1.4 Objective*

The objective of this chapter is to use symptoms and signs of severe illness documented at the time of admission to a large urban neonatal unit as predictors to:

i. Develop and internally validate a model to predict in-hospital mortality
ii. Externally validate the model and compare performance against NETS

## 6.2 Methods

The methods applied in this Chapter are for the most part the same as described in Chapter 5. These encompass model development (derivation) and validation (internal and external). In the methods section I will therefore refer to Chapter 5 where applicable and focus on describing aspects that are specific to this Chapter.

### *6.2.1 Model derivation*

*a) Source of data and study participants*

I used the same data set described in Chapter 5, Section 5.2.1 (b) comprising of 5427 participants. However, since the treatment sheets were not relevant to the SENSS model I did not exclude the 587 participants with missing treatment sheets.

*b) Outcome*

This remains in-hospital mortality as defined in Chapter 5.

*c) Predictors*

In Table 6.1, I presented eight candidate predictors. I also considered floppiness or inability to suck and bulging fontanelle for a total of ten candidate predictors. Exploratory analysis of the study data set revealed that eight predictors were available: difficulty feeding, convulsions, indrawing, temperature, respiratory rate, central cyanosis, floppy/inability to suck, bulging fontanelle. I excluded temperature and respiratory rate as these were poorly documented (89% and 65% missingness respectively) leaving six candidate predictors shown in Table 6.2.

***Table 6.2**: Symptoms and signs of severe illness considered as candidate predictors*

| Variable name | Description | Definition | |
|---|---|---|---|
| | | Yes | No* |
| **Difficulty feeding** | Symptom | Present | Absent |
| **Convulsions** | Symptom | Present | Absent |
| **Indrawing** | Sign | Present | Absent |
| **Central cyanosis** | Sign | Present | Absent |
| **Floppy/inability to suck** | Sign | Present | Absent |
| **Bulging fontanelle** | Sign | Present | Absent |

*Documented that the patient did not have the symptom or sign present

These data were extracted from the NAR as documented by the admitting clinician. For each of these predictors, the admitting clinician ticks on the NAR whether it is present or absent based on their assessment of the neonate. I also excluded bulging fontanelle from the list as it was a rare sign (present in 0.2% of the derivation data set) as shown in Table 6.6. This left 5 predictors to which I added birthweight and sex (as defined in Table 5.3) for a total of seven.

*d) Sample size*

I used the events per variable (EPV) criteria to assess the adequacy of the sample size in the derivation data set. I had seven predictors for the SENSS model but dummy variables were created during logistic regression giving a total of 10 parameters. There were thus 10 parameters for the 445 in-hospital deaths observed yielding an EPV of 45 (445/10).

*e) Missing data*

Besides predictors, the handling of missing data was another key difference between the NETS and SENSS models. In this section I describe how I implemented multiple imputation to fill in missing data with plausible values. First, I will describe the pattern of missingness of the predictors. I will then explain the multiple imputation procedures and show the results. Table 6.3 shows the distribution and pattern of missingness of the predictors.

***Table 6.3:*** *Distribution of predictors and missingness in the SENSS derivation dataset*

| Characteristic | Category | N= 5427 | |
|---|---|---|---|
| | | n | % |
| **1.Sex** | | | |
| | Male | 2937 | 54 |
| | Missing | 13 | 0.2 |
| **2.Birthweight(kg)** | | | |
| | <1 | 32 | 0.5 |
| | 1-<1.5 | 136 | 3 |
| | 1.5-<2.5 | 1180 | 22 |
| | 2.5-4 | 3841 | 71 |
| | >4 | 229 | 4 |
| | Missing | 9 | 0.2 |
| **3.Difficulty feeding** | | | |
| | Yes | 225 | 5 |
| | Missing | 575 | 11 |
| **4.Convulsions** | | | |
| | Yes | 29 | 0.5 |
| | Missing | 571 | 11 |
| **5.Central cyanosis** | | | |
| | Yes | 222 | 5 |
| | Missing | 570 | 11 |
| **6.Severe indrawing** | | | |
| | Yes | 429 | 9 |
| | Missing | 620 | 11 |
| **7.Floppy/unable to suck** | | | |
| | Yes | 530 | 11 |
| | Missing | 876 | 16 |
| **8.Outcome** | | | |
| | Dead | 508 | 9 |
| | Missing | 19 | 0.4 |

Missing values arose in the dataset where the admitting clinician failed to document the presence or absence of the clinical predictor in the patient record. The four-

step data quality assurance mechanism described in Chapter 2 (Section 2.3.3) ensures that missing values are only due to failure of documentation rather than data entry errors.

The proportion of predictor missingness (also termed variable wise missingness) ranged from 0.2% (sex and birthweight) to 16% (floppy/unable to suck). Looking at the signs and symptoms only missingness ranged from 11% (difficulty in breathing, convulsions, central cyanosis, severe indrawing) to 16% (floppy/unable to suck). I adopted the missing at random (MAR) assumption as described in Chapter 2 (Section 2.6.2 (c)) as the explanatory mechanism for the observed missingness.

I specified an imputation model to include the seven predictors, outcome and auxiliary variables as recommended by Ewout Steyerberg.(136) Auxiliary variables are those that are related to the predictors but are not included in the prediction model. There is empirical evidence supporting inclusion of auxiliary variables as increasing the number of variables in the imputation model makes the MAR assumption more plausible.(136) I chose the key clinical characteristics described in Chapter 4 (Table 4.2): mode of delivery, place of delivery, HIV exposure status and documentation of Apgar score as the auxiliary variables. Examination of the dataset for case wise missingness revealed that 31% of individuals in the dataset had at least one of either predictors, auxiliary variables or outcome missing. I therefore used this as a guide to setting the number of imputation models at 31, based on published recommendations.(289)

Multiple imputation using the chained equation approach was implemented using the “aregImpute” function of the rms package in R.(290) The R code is included in Appendix 9. Missing values for binary variables were imputed using logistic regression while categorical variables were imputed with polytomous logistic regression. There were no continuous variables. A comparison of the distribution of the SENSS predictors in the original dataset and in one of the multiply imputed data sets is shown in Table 6.5 of the results section.

*f) Statistical analysis methods*

I derived the SENSS model using the 31 imputed data sets by applying Rubin’s Rules.(291) This entailed specifying a logistic regression model to predict in-hospital mortality as described in Chapter 5 (Section 5.2.1 (g) for each of the 31 data sets. The 31 sets of parameter estimates were thereafter combined to obtain a single set of regression coefficients and performance estimates. The key parameters of the model’s apparent performance were discrimination (c-statistic) and calibration (calibration slope) and were computed as described in Chapter 5 using the val.prob.ci.2 function.(204) Multiple imputation analyses were implemented using the “fit.mult.impute” function of the rms package in R and are included in Appendix 9.(290)

g) Risk groups

I did not create risk groups at this stage for the same reason described in Chapter 5 (Section 5.2.1 (h))

*6.2.2 Internal validation*

Internal validation was carried out using a bootstrap resampling procedure as described in Chapter 5 (Section 5.2.2)

*6.2.3 External Validation*

*a) Source of data*

This was the same study hospital as described in Chapter 5 (Section 6.2.1 (a)) and Chapter 2 (Section 2.3).

b) Participants

The same data set described in Chapter 5 (Section 5.2.3 (b) and Figure 5.2) was used. However, I did not exclude observations with missing treatment sheets as this did not affect the SENSS predictors. As a result, the number of participants remained as 1627. I provide a detailed description of the handling of missing data in Section 6.2.3 (f)).

*c) Outcome and Predictors*

The same definitions of outcome and predictors described for model derivation (Section 6.2.1 (c) and (d)) were used.

*d) Sample size*

The external validation data set contains 151 events (in-hospital deaths) and 1476 non-events (no deaths). This is more than the minimum recommended 100 events and 100 non-events as described in Chapter 5 (Section 5.2.3 (d)).

*e) Missing data*

Table 6.4 shows the distribution of the predictors and the pattern of variable wise missingness in the external validation dataset.

***Table 6.4**: Distribution of predictors and missingness in the external validation dataset*

| Characteristic | Category | N*= 1627 | |
|---|---|---|---|
| | | n | % |
| **1.Sex** | | | |
| | Male | 961 | 58 |
| | Missing | 2 | 0.1 |
| **2.Birthweight(kg)** | | | |
| | <1 | 10 | 0.6 |
| | 1-<1.5 | 45 | 3 |
| | 1.5-<2.5 | 361 | 22 |
| | 2.5-4 | 1125 | 69 |
| | >4 | 85 | 5 |
| | Missing | 9 | 0.2 |
| **3.Difficulty feeding** | | | |
| | Yes | 134 | 9 |
| | Missing | 199 | 12 |
| **4.Convulsions** | | | |
| | Yes | 25 | 2 |
| | Missing | 199 | 12 |
| **5.Central cyanosis** | | | |
| | Yes | 64 | 5 |
| | Missing | 216 | 13 |
| **6.Severe indrawing** | | | |
| | Yes | 62 | 5 |
| | Missing | 234 | 14 |
| **7.Floppy/unable to suck** | | | |
| | Yes | 359 | 26 |
| | Missing | 222 | 13 |
| **8.Outcome** | | | |
| | Dead | 151 | 9 |
| | Missing | 6 | 0.4 |

***** Total number of patients in the data set

The proportion of predictor missingness ranged from 0.1% (sex) to 14% (severe indrawing). Looking at the signs and symptoms only, missingness ranged from 12% (difficulty in breathing and convulsions) to 14% (severe indrawing). I conducted multiple imputation for missing values under the MAR assumption using the same procedures described for model derivation (Section 6.2.1 (e)). The only

difference was in the number of imputed data sets. Examination of case wise missingness in the external validation data set demonstrated that at least 23% of the individuals in the dataset had at least one missing value for the variables included in the imputation model. I therefore generated 23 imputed data sets.

*f) Statistical analysis methods*

Analyses for external validation were conducted using the same methods I described in Chapter 5 (section 5.2.3 (f)). The main steps were the construction of an initial calibration curve to determine the need for model updating. The model was thereafter updated by adjusting the intercept thereby creating the final SENSS model.

## 6.3 Results

### *6.3.1 Description of the multiply imputed data sets*

Table 6.5 compares the distribution of the predictors and outcome between the original and imputed data sets in the derivation and external validation data sets.

***Table 6.5**: Distribution of predictors in the original and imputed data sets*

| Characteristic | Category | Derivation (N*=5427) | | | | External validation (N*=1627) | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Original | | Imputed | | Original | | Imputed | |
| | | n | % | n | % | n | % | n | % |
| **1.Sex** | | | | | | | | | |
| | Male | 2937 | **54** | 2942 | **54** | 961 | **58** | 962 | **59** |
| **2.Birthweight(kg)** | | | | | | | | | |
| | <1 | 32 | **0.5** | 32 | **0.6** | 10 | **0.6** | 10 | **0.6** |
| | 1-<1.5 | 136 | **3** | 136 | **3** | 45 | **3** | 45 | **3** |
| | 1.5-<2.5 | 1180 | **22** | 1182 | **22** | 361 | **22** | 361 | **22** |
| | 2.5-4 | 3841 | **71** | 3848 | **71** | 1125 | **69** | 1126 | **69** |
| | >4 | 229 | **4** | 229 | **4** | 85 | **5** | 85 | **5** |
| **3.Difficulty feeding** | | | | | | | | | |
| | Yes | 225 | **5** | 264 | **5** | 134 | **9** | 146 | **9** |
| **4.Convulsions** | | | | | | | | | |
| | Yes | 29 | **0.5** | 38 | **0.5** | 25 | **2** | 29 | **2** |
| **5.Central cyanosis** | | | | | | | | | |
| | Yes | 222 | **5** | 255 | **5** | 64 | **5** | 79 | **5** |
| **6.Severe indrawing** | | | | | | | | | |
| | Yes | 429 | **9** | 494 | **9** | 62 | **5** | 69 | **4** |
| **7.Floppy/unable to suck** | | | | | | | | | |
| | Yes | 530 | **11** | 647 | **11** | 359 | **26** | 413 | **25** |
| **8.Outcome** | | | | | | | | | |
| | Dead | 508 | **9** | 509 | **9** | 151 | **9** | 151 | **9** |

***** Total number of patients in the data set

The proportions observed for each variable in the imputed data sets demonstrate that the multiple imputation procedures filled in the missing values while preserving the pattern of distribution observed in the original data sets. In Table 6.6, I focus on the imputed datasets only as these are the data used for modelling.

Table 6.6 demonstrates that there was a temporal difference in the distribution of the signs and symptoms used as predictors in the SENSS model amongst patients who died. For four out of the five predictors, the proportion of observations amongst the patients who died is higher in the derivation validation data set in comparison to the external validation data set.

***Table 6.6**: Distribution of predictors by in-hospital mortality in the imputed datasets*

| Predictor | Derivation (n=5427) * | | | | External validation (n=1627) † | | | |
|---|---|---|---|---|---|---|---|---|
| | Present | | Dead‡ | | Present | | Dead‡ | |
| | n | % | n | % | n | % | n | % |
| **1.Difficulty feeding** | 264 | **5** | 110 | **42** | 146 | **9** | 42 | **29** |
| **2.Convulsions** | 38 | **1** | 17 | **45** | 29 | **2** | 7 | **24** |
| **3.Central cyanosis** | 255 | **5** | 141 | **56** | 79 | **5** | 35 | **44** |
| **4.Severe indrawing** | 494 | **9** | 266 | **54** | 69 | **4** | 40 | **59** |
| **5.Floppy/unable to suck** | 647 | **12** | 297 | **46** | 413 | **25** | 127 | **31** |

* One of the 31 multiply imputed data sets
† One of the 23 multiply imputed data sets
‡ Number and proportion dead amongst those with the predictor present

The largest difference is for convulsions with this symptom observed in 45% of the patients who died in the derivation data set compared to 24% in the external validation data set. The smallest difference is in severe indrawing with 54% versus 59% for the derivation and external validation datasets respectively.

*6.3.2 Model derivation*

*a) Model specification*

The logistic regression model specified to predict in-hospital mortality is shown in Box 6.1 and the detailed code is included in Appendix 9.

```
fit.mult.impute(outcome~birthweight + sex + difficulty feeding +
convulsions + indrawing + central cyanosis +floppy/unable to suck,
lrm,xtrans=senssdata,data=missdata,x=T,y=T,linear.predictors = T)
```

***Box 6.1***: *Logistic regression model for the SENSS score*

In contrast to the NETS model, there were no missing variables (after multiple imputation) and therefore no observations dropped by complete case analysis.

*b) Model estimation*

The coefficients estimated by logistic regression were averaged over the 31 imputed data sets giving the results shown in Table 6.7.

***Table 6.7**: Coefficients from the logistic regression model for the NETS score*

| Model parameter | Coefficients* | Standard Error | P value |
|---|---|---|---|
| **Intercept** | -3.858 | 0.130 | <0.0001 |
| **ELBW(<1kg)** | 5.758 | 1.056 | <0.0001 |
| **VLBW(1-<1.5kg)** | 3.708 | 0.290 | <0.0001 |
| **LBW(1.5-<2.5kg)** | 0.923 | 0.141 | <0.0001 |
| **Macrosomia(>4kg)** | -0.492 | 0.457 | 0.2821 |
| **Male** | -0.134 | 0.130 | 0.3042 |
| **Difficulty feeding** | 1.360 | 0.210 | <0.0001 |
| **Convulsion** | 1.398 | 0.480 | 0.0035 |
| **Severe indrawing** | 1.979 | 0.167 | <0.0001 |
| **Central cyanosis** | 0.958 | 0.215 | <0.0001 |
| **Floppy/ unable to suck** | 1.627 | 0.166 | <0.0001 |

* Values in log odds scale for the predictors with in-hospital mortality as the outcome
ELBW: Extremely Low Birth Weight, VLBW: Very Low Birth Weight, LBW: Low Birth Weight

The coefficients are in the log odds scale as described in Chapter 5 (Section 5.3.2 (c)).

*c) Model apparent performance*

*Overall performance*

Overall performance metrics for the SENSS model are displayed in Table 6.8. As was the case for the NETS model, I focused on measures of discrimination and calibration. I report these results in the next section.

***Table 6.8**: Overall measures of NETS model apparent performance*

| Index | Result | Scale | Interpretation |
|---|---|---|---|
| **Nagelkerke $R^2$** | 0.508 | 0 to 1 | Higher when those without outcome are closer to 0% predicted probability and those with closer to 100% |
| **Brier score** | 0.048 | 0 to 1 | 0 perfect discrimination |

*Discrimination and calibration*

Figure 6.1 below is the calibration plot for the SENSS model at derivation stage. This is the model's apparent performance as described in Chapter 2.

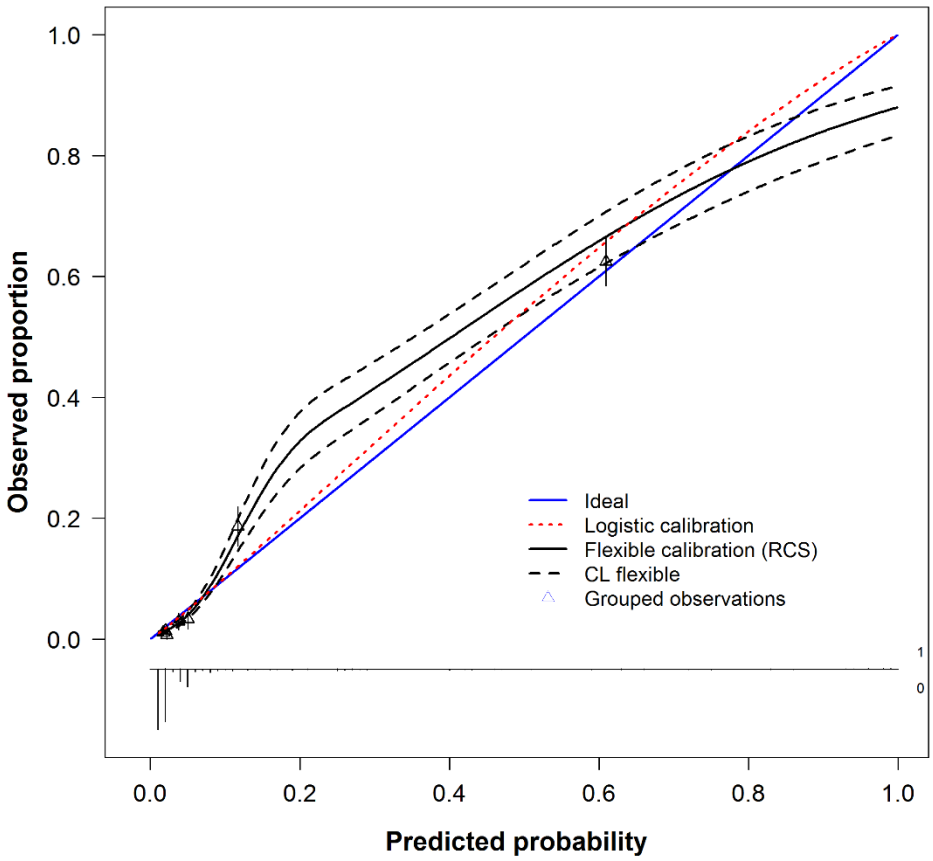



***Figure 6.1**: Calibration plot for SENSS in the derivation data set*

The SENSS model discrimination as measured by the c-statistic of the logistic regression model in the derivation data set is 0.91 (95% C. I 0.89 to 0.93). This means that in the derivation data set, the predicted probability of in-hospital death for patients who died (observed death) was higher than the predicted probability in those who left the unit alive 91% of the time.

As discussed in Chapter 5 (Section 5.3.2 (d)) regression models always give an intercept of zero and slope of one (perfect calibration) for the derivation data.(135) This reflects the modelling process where the predicted odds of death (and hence the probability of death) given the data are perfectly matched to the observed and thus calibration is always perfect. This is however not the case at internal validation the results of which I present in the next section.

### *6.3.3 Internal validation*

#### *a) Correction of model performance for optimism*

Table 6.9 contains the optimism corrected overall performance measures for the SENSS model. There is little difference between the original indices and the corrected indices.

***Table 6.9**: Performance of the SENSS model after internal validation*

| Index | Original index | Training | Test | Optimism | Corrected index |
|---|---|---|---|---|---|
| **Nagelkerke $R^2$** | 0.502 | 0.515 | 0.507 | 0.007 | 0.494 |
| **Intercept*** | 0.000 | 0.000 | -0.029 | 0.029 | -0.029 |
| **Slope*** | 1.000 | 1.000 | 0.986 | 0.014 | 0.986 |

* Calibration intercept and slope

As I described in Chapter 5 (Section 5.3.3) this shows that there was little overfitting during model derivation because of the large sample size used, limited number of candidate variables and no variable selection procedures. Of interest is the calibration intercept and slope which remain very close to 0 and 1. This effect of the large sample size is also illustrated in Table 6.10 which shows there was little change in the c-statistic.

***Table 6.10**: Discrimination of the SENSS model after correction for optimism*

| Original | Training | Test | Optimism | Corrected |
|---|---|---|---|---|
| **0.902** | 0.901 | 0.899 | 0.001 | 0.901 |

Correction for optimism only slightly reduced the c-statistic from 0.902 (original) to 0.901(corrected). Similar results were obtained after shrinkage of the model coefficients which I present in the next section.

*b) Shrinkage of the model coefficients after internal validation*

As was the case with the NETS model, there was little change in the SENSS model coefficients after shrinkage to adjust for overfitting. This is shown in Table 6.11.

***Table 6.11**: Shrinkage of the SENSS model coefficients to correct for optimism*

| Model parameter | Original* | Shrunk* |
|---|---|---|
| **Intercept** | -3.858 | -3.804 |
| **ELBW(<1kg)** | 5.758 | 5.677 |
| **VLBW(1-<1.5kg)** | 3.708 | 3.656 |
| **LBW(1.5-<2.5kg)** | 0.923 | 0.910 |
| **Macrosomia(>4kg)** | -0.492 | -0.485 |
| **Male** | -0.134 | -0.132 |
| **Difficulty feeding** | 1.360 | 1.341 |
| **Convulsion** | 1.398 | 1.378 |
| **Severe indrawing** | 1.979 | 1.951 |
| **Central cyanosis** | 0.958 | 0.945 |
| **Floppy/ unable to suck** | 1.627 | 1.604 |

* Values in log odds scale for the predictors with in-hospital mortality as the outcome
ELBW: Extremely Low Birth Weight, VLBW: Very Low Birth Weight, LBW: Low Birth Weight

Given these results of internal validation, I concluded that there was insignificant optimism in the original model and I therefore carried it forward to external validation without updating the coefficients or intercept. I present the results of the external validation in the next section.

*6.3.4 External validation*

*a) Model performance*

The results of the temporal external validation of the SENSS model are displayed in Figure 6.2.

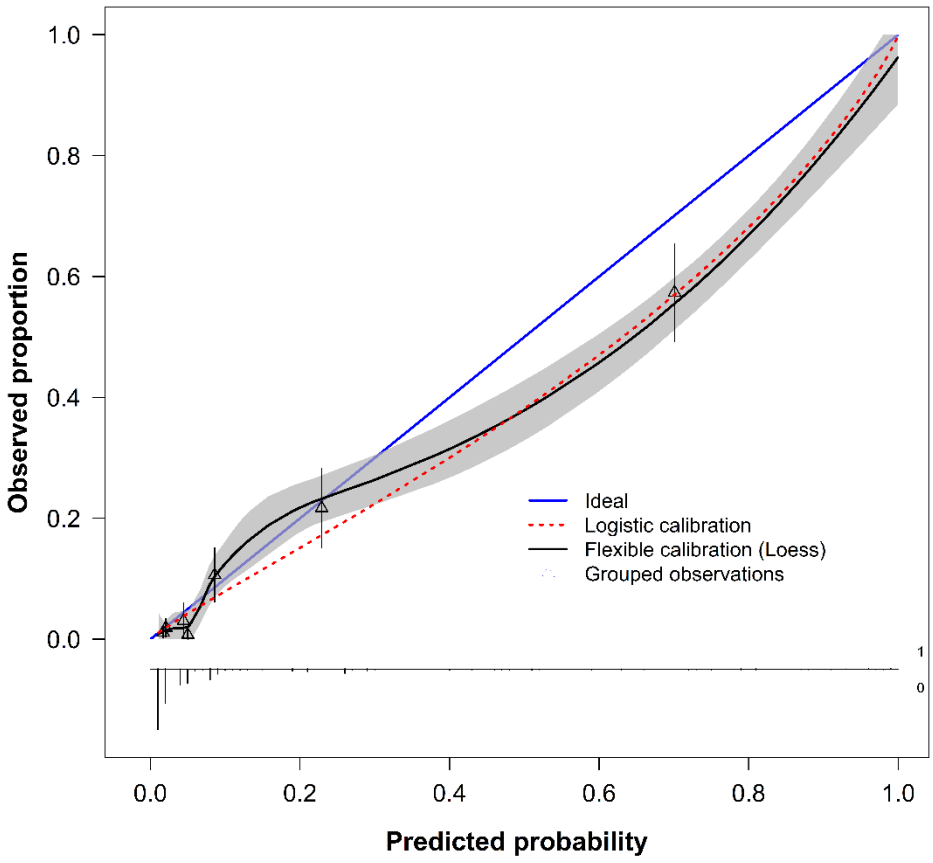



***Figure 6.2**: Calibration curve for SENSS in the external validation data set*

The point estimate of the c-statistic declined from 0.92 in the derivation data set to 0.89 (95% C.I 0.84 to 0.93) in the external validation data set. Whereas in the derivation data set calibration was perfect (intercept 0, slope 1), Figure 6.2 shows the extent of mis-calibration in the new population. The new intercept is -0.33 (95% C.I -0.56 to -0.11) and slope 0.90 (95% C.I 0.78 to 1.01). This means that the predicted in-hospital probability of death is higher than the observed across the

range of predicted risk. This miscalibration is greater at predicted probabilities greater than 0.2 where the SENSS predictions (represented by the logistic and flexible calibration curves) lies below the ideal (perfect) calibration line showing over-prediction. To correct this miscalibration, I updated the model intercept and present the updated model in the next section.

*b) Model updating after external validation*

I used the same procedures to update the SENSS model intercept that I described in Chapter 5 (Section 5.2.3 (f)). I computed a correction factor which I then added to the original intercept to give the updated intercept. The updated intercept is -4.1946 (original -3.8583) while the predictor coefficients remain the same as shown in Table 6.12.

***Table 6.12**: Coefficients of the logistic regression model for the updated SENSS score*

| Model parameter | Coefficients* |
|---|---|
| **Intercept** | -4.1946 |
| **ELBW(<1kg)** | 5.7580 |
| **VLBW(1-<1.5kg)** | 3.7082 |
| **LBW(1.5-<2.5kg)** | 0.9232 |
| **Macrosomia(>4kg)** | -0.4918 |
| **Male** | -0.1336 |
| **Difficulty feeding** | 1.3596 |
| **Convulsion** | 1.3977 |
| **Severe indrawing** | 1.9790 |
| **Central cyanosis** | 0.9584 |
| **Floppy/ unable to suck** | 1.6266 |

* Values in log odds scale for the predictors with in-hospital mortality as the outcome
ELBW: Extremely Low Birth Weight, VLBW: Very Low Birth Weight, LBW: Low Birth Weight

I present the performance of this final model next.

*c) Performance of final model*

Adjustment of the original model intercept reduces the miscalibration observed at external validation. The logistic regression formula for the final model is shown in Box 6.2.

```
Linear predictor =-4.1946 + 5.7580*ELBW +3.7082*VLBW + 0.9232*LBW –
                  0.4918*Macrosomia - 0.13596*Male + 1.3596*Feeds +
                  1.3977*Convulsion - 1.9790*Indrawing + 0.9584*Cyanosis +
                  1.6266*Floppy/unable to suck
```

***Box 6.2**: Final logistic regression model for SENSS after updating the intercept*

Box 6.2 shows that the original predictor coefficients obtained at model derivation (Table 6.7) have been retained in the final model. The performance of the SENSS final model is illustrated in the calibration plot in Figure 6.3. Model discrimination (c-statistic) remains the same as in the original 0.89 (95% C.I 0.84 to 0.93).

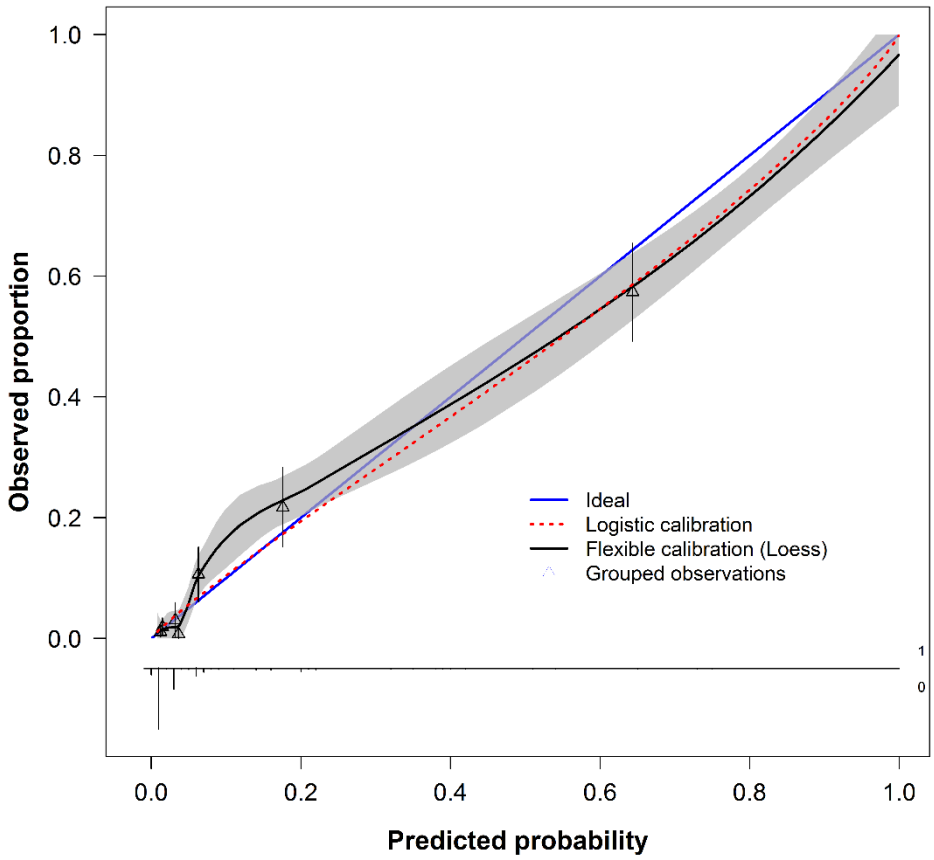



***Figure 6.3**: Calibration curve for the final SENSS model update of the intercept*

In Figure 6.3 we observe that there is now less over prediction of probability of in-hospital death at higher (>0.4) predicted probabilities. The new intercept is 0 (95% C.I -0.22 to 0.22) while the slope remains 0.90 (95% C.I 0.78 to 1.01). Adjustment of the intercept shifts the calibration curve (both logistic and flexible) upwards to a position closer to (but still below) the ideal calibration line in comparison to its position in Figure 6.2. This however comes at a cost as there is now greater under prediction at lower (<0.4) probabilities given the position of the model's curve above the ideal line.

## 6.4 Discussion

### *6.4.1 Summary of findings*

In this Chapter, I have demonstrated the use of five symptoms and signs (clinical predictors) of severe illness in addition to birthweight and sex to predict the probability of in-hospital death. This Score for Essential Neonatal Symptoms and Signs (SENSS) thus combines clinical predictors currently used individually as predictors of neonatal severe illness into a multivariable prognostic model. The five clinical predictors are two symptoms (history of difficulty feeding, history of convulsions) and three signs (severe indrawing, central cyanosis and floppiness/inability to suck). These predictors were pre-selected based on their current use in clinical practice which is informed by evidence from a systematic review of predictors of severe illness in neonates and young infants.(17, 51)

Using complete case analysis to derive the SENSS model would have resulted in the dropping of 1046/5427 (19%) of the observations in contrast to NETS model where 40/4840 (0.8%) were excluded. Whereas 587/5427 (11%) of observations were excluded from the NETS model due to missing treatment sheets, the pattern of missingness was missing completely at random (Table 5.7, Chapter 5) and in any case the total number excluded from NETS 627/5427 (12%) is still less than the 19% for SENSS. I therefore used multiple imputation to fill in the missing values before model estimation. Just like in Chapter 5 (NETS model), I used the full model approach (no selection of variables during model estimation).

The aim of this Chapter is to conduct a head to head comparison of the performance of the NETS model and an alternative model using predictors

potentially available from routine data in settings providing essential neonatal care. I therefore first summarise the performance of SENSS in Table 6.13 before comparing its development and predictive performance with NETS in the subsequent sections.

***Table 6.13:*** *Summary of the predictive accuracy of SENSS*

| Parameter | Metric | Point estimates by modelling stage | | |
|---|---|---|---|---|
| | | **Derivation** | **Internal validation*** | **External validation** |
| **1. Discrimination** | c-statistic | 0.91 | 0.90 | 0.89† |
| **2. Calibration** | intercept | 0 | -0.03 | 0† |
| | slope | 1 | 0.99 | 0.90† |

* Optimism corrected performance
† Final model after updating

There was good discriminatory performance as measured by the c-statistic at derivation, internal validation and external validation. The c-statistic was above the recommended 0.80 at the three stages. As I discussed in Chapter 5, the calibration intercept and slope are always 0 and 1 for regression models at derivation and thus are not useful performance measures at this stage. Of interest, is calibration at internal validation and more so at external validation. Given the large data set used at derivation with 509 deaths for 10 model parameters (EPV=50) there was little optimism in the SENSS model. This is reflected in the minor change observed in discrimination (c-statistic) and calibration (slope and intercept) after correction for optimism through internal validation.

*6.4.2 Predictor selection*

I used subject knowledge to select predictors for the NETS and SENSS models (Section 5.1.2 and Section 6.1.2 respectively). I thereafter used the full model approach to model estimation as described in Section 5.2.1 and Section 6.2.1 respectively. Both models include birthweight in the five commonly used clinical categories. As described in Chapter 4, multivariable modelling is one way to address the residual variation in these birthweight categories. In exploratory analyses not included in this Thesis, I explored models with birthweight as the only predictor specified alternatively as continuous, binary and categorical (five or six categories). All these models had discrimination by c-statistic less than the recommended 0.8 in comparison to greater than 0.8 for NETS and SENSS.(153)

Predictor selection after specification of the full model was not done for the following reasons:

i. The ideal case in prediction modelling is a limited number of preselected predictors entered into a full model. (96, 136, 292) In this regard, the NETS and SENSS models are based on a limited set of preselected predictors: five essential treatments for NETS and five symptoms and signs drawn from a systematic review for SENSS.(17, 51) As described in Chapter 5 (Section 5.1.2) and Chapter 6 (Section 6.1.2), all these predictors represent clinically important constructs.(136, 292)
ii. A more parsimonious model (fewer predictors) may be expected to have better predictive performance.(136) However apart from the 0.8 threshold for good discrimination there is no consensus on thresholds of relative

discriminatory ability or difference in calibration when comparing alternative models.(153)

I would have considered predictor selection (in this case backward selection since I started with full models) under the following two scenarios (136, 292):

i. If there were a large set of predictors particularly if these are likely to be unavailable for future patients and/or are costly to collect.(136) This was not the case as NETS and SENSS include seven predictors all drawn from routine data and none requires diagnostic testing or electronic monitoring.
ii. If the predictors demonstrated small or implausible effects after model estimation.(136) However, all the predictors included demonstrated clinically plausible effects. In addition, exclusion based on statistical significance is not recommended and the inclusion of such “noise” variables may reduce predictive performance only marginally particularly where there are few candidate predictors such as in NETS and SENSS.(136, 292)

However, there is no consensus yet on the best predictor selection strategy. (3, 4, 7, 8). In addition, backward selection has potential disadvantages. First is instability of the selection where selected predictors may vary depending on the characteristics of the data used.(96, 136, 292, 293) Secondly, the coefficients which represent the contribution of each predictor to the outcome probability may be erroneously large.(136, 292) Thirdly, the variability of the predictor coefficients may be misspecified as the respective standard errors are computed as if each model was pre-specified.(136) Erroneous coefficients and standard errors will in

turn result in exaggerated p-values.(136) Finally, there is evidence that stepwise selected models perform worse at internal and external validation. (136)

Nevertheless, predictor selection may in future become increasingly important for prediction modelling in low resource settings. This is because as in-hospital care for neonates develops e.g. new laboratory tests and therapeutics like BCPAP new candidate predictors shall be available. Selection may then be required for model updating (NETS or SENSS) or in developing new models. This may involve the use of newer methods such as Lasso (least absolute shrinkage and selection operator), bootstrapping for selection, Bayesian Model Averaging and Bagging and boosting.(136) However, it remains unclear whether these methods eliminate the disadvantages of stepwise selection.(136)

Another approach to predictor selection is reclassification. This includes reclassification tables, reclassification calibration, Net Reclassification Index and Integrated Discrimination Index.(277) These measure how many subjects are reclassified to the appropriate outcome status as observed by addition of a new predictor.(277) In the case of SENSS and NETS this would involve checking which of the included predictors results in appropriate reclassification. The most popular of these methods is the Net Reclassification Index (NRI) which compares two models by computing the net change in right classification.(136, 294) However the statistical validity of NRI is yet to be established and therefore I did not consider it further.(295-298)

*6.4.3 Comparison of performance between NETS and SENSS at external validation*

The most useful aspect of prediction model performance is the performance at external validation.(106) By evaluating the model in an external population an assessment of its generalisability and potential clinical utility is made. In this thesis, the population of interest is neonates admitted to neonatal units that offer essential neonatal interventions in low resource settings. I conducted a temporal external validation of SENSS and NETS in the same data set. For both models, there was observed deterioration in performance and I therefore adjusted the intercept of each model using the same procedure.

After this update, discrimination was similar for NETS and SENSS with c-statistic of 0.89 (95% C.I 0.86 to 0.92) and 0.89 (95% C.I 0.84 to 0.93) respectively. Figure 6.4 compares the calibration curves for the two models. The point estimates for the calibration intercepts were both zero with similar confidence intervals: 0.00 (95% C.I -0.24 to 0.24) for NETS and 0.00 (95% C.I -0.22 to 0.22) for SENSS. This similarity in the intercept is expected as the statistical method used to update the intercept for both models fixes the point estimate at zero.

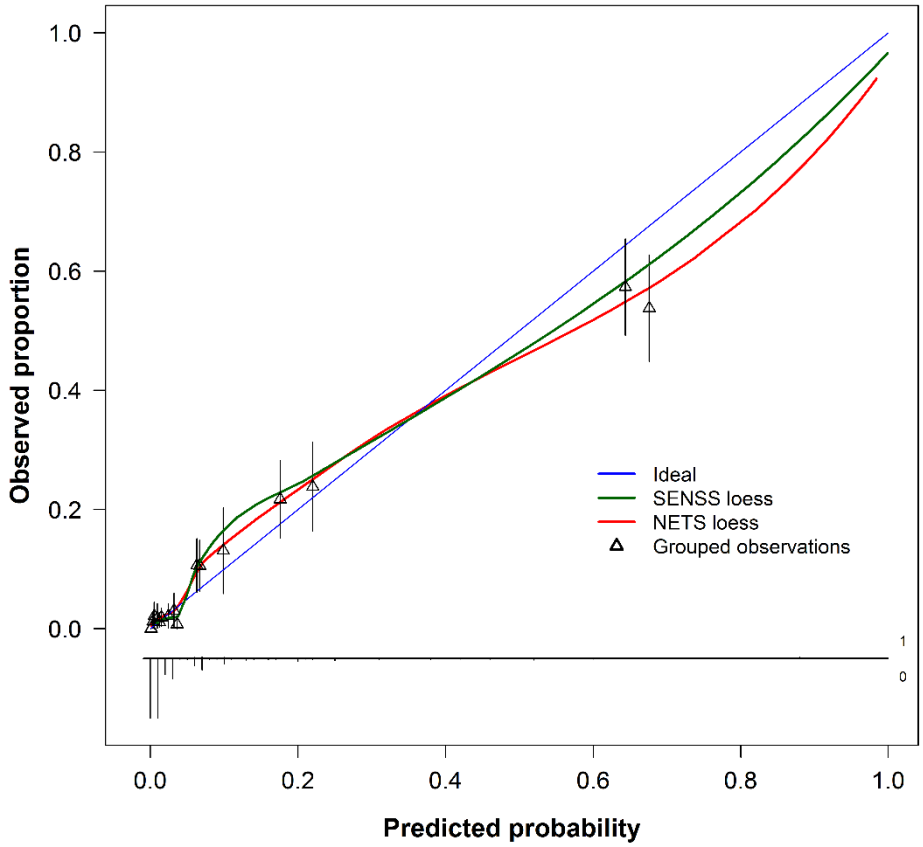



***Figure 6.4**: Performance of NETS and SENSS at external validation in the same population*

In the next section I explore the possible reasons for the observed difference in the calibration slopes.

### *6.4.4 Potential reasons for difference in calibration between NETS and SENSS*

The calibration slope compares the agreement between the predicted and the observed risk across the range of predicted risks for the population. A perfectly calibrated model has a slope of one. This means the closer the slope is to one the better the agreement between the observed and the predicted risk. The final SENSS model had a better calibration slope (0.90, 95% C.I 0.78 to 1.01) than the

NETS model (0.76, 95% C.I 0.65 to 0.87). This means that across the range of predicted risks, the SENSS predictions were closer to the observed outcomes in the external validation population.

Further inferences can be made by examining the performance of the two models by the tenths of predicted risks represented by the triangles on Figure 6.4. At lower (<0.4) predicted risks the curves for both models largely lie above but close to the ideal (slope=1, perfect calibration slope). This means that across this range of predicted risks both models slightly under predict in-hospital death. However, at higher (>0.4) predicted risks, both models slightly over predict risk of in-hospital death given the position of the calibration curves beneath the ideal line.

In general, there are three reasons why the performance of prediction models deteriorates at external validation as described in Chapter 5. These are: (a) differences in definition and measurement of predictors and outcomes (b) differences in the patients included (case-mix) and (c) the inclusion of fewer patients in validation compared to derivation.(280) I will now explore each of these factors as potential reasons that could explain the better calibration of SENSS compared to NETS model.

a) Differences in definition and measurement of predictors

Given that the same data collection system was used for the derivation and validation data sets there was no change in the definitions of the predictors with regards to data entry. However, the documentation of the predictors in the patient file is dependent upon the admitting clinician. The admitting clinicians often change

for example due to rotations between departments, transfers and recruitment of new staff. Prescription of treatments (NETS predictors) and recognition of signs and symptoms of severe illness (SENSS predictors) are different clinical skills. Predictor definition and measurement at clinician level is dependent on these clinical skills. It is therefore conceivable that there was a systematic difference in the effect of secular trend on these two constructs at clinician level contributing to the difference in model performance.

b) Differences in case-mix

Whereas external validation was done with same data set, there may be differences in the distribution of predictors of the two models by the outcome (in-hospital mortality). Table 6.14 summarises the distribution of treatments (NETS) and signs and symptoms (SENSS) by in-hospital mortality in the derivation and external validation data sets. This table shows the proportion of patients with each of the predictors present who experienced the predicted outcome (in-hospital). In general, the proportion of deaths were less in the external validation dataset. However, the magnitude of these differences vary between NETS and SENSS datasets which may contribute to the differences in model calibration observed at external validation.

***Table 6.14**: Distribution of predictors by in-hospital mortality for NETS and SENSS*

| Model and predictors | Derivation | External Validation |
|---|---|---|
| | % Dead | % Dead |
| **1.SENSS*** | | |
| **Difficulty feeding** | 42 | 29 |
| **Convulsions** | 45 | 24 |
| **Central cyanosis** | 56 | 44 |
| **Severe indrawing** | 54 | 59 |
| **Floppy/unable to suck** | 46 | 31 |
| **2.NETS†** | | |
| **Intravenous antibiotics** | 20 | 16 |
| **Intravenous fluids** | 34 | 23 |
| **Enteral feeds** | 1 | 1 |
| **Oxygen** | 41 | 24 |
| **Phenobarbital** | 61 | 25 |

*SENSS: Score for Essential Neonatal Symptoms and Signs
† NETS: Neonatal Essential Treatment Score

### *6.4.5 Justification for external validation*

The NETS and SENSS models both demonstrate minimal optimism after internal validation (Table 5.9, Table 5.10 and Table 6.10, Table 6.11 respectively). This means that the models are likely to perform well in the study population without requiring external validation. However, the temporal external validation conducted provides useful insights.

To begin with, as noted in Section 6.4.4, smaller sample size at external validation is one of the reasons deterioration in performance may be observed. This was indeed the case for NETS and SENSS but the sample size was still adequate for external validation as there were more than 100 events (deaths) and 100 non-events (alive).(136) Arguably therefore, the deterioration in performance could be a result of real differences in case-mix and clinical practices over time (Section 6.4.4). In addition, all prognostic models require updating over time (further discussed in Chapter 8) and for external populations the NETS and SENSS models will require geographical external validation.(136, 147) The external validation of

NETS and SENSS thus demonstrates how model updating for the index hospital and geographical external validation for other hospitals respectively may be conducted.

### *6.4.6 Limitations*

The main limitation for the SENSS model is the missingness of the predictors observed in the data sets. In addition, clinicians may have inconsistent definition of signs and symptoms besides inter-observer variation in the definition of signs. Application of the model in routine practice will therefore require better documentation of these clinical symptoms and signs given that 19% of the observations had missing data at the derivation stage which would have been dropped if a complete case analysis approach was used. One way of improving documentation in routine practice is through audit and feedback linked to standard record forms such as the NAR as is currently underway in several hospitals in Kenya.(281, 283, 288) As is the case for the NETS model, improvement in survival will reduce the usefulness of in-hospital mortality as an outcome. However, as I described in Chapter 5 a decrease in mortality is likely to be accompanied by an increase in morbidity and the model may then be updated to predict for example a composite outcome that includes death and severe in-hospital morbidity.

### *6.4.7 Implications*

The SENSS model demonstrated better calibration than the NETS model in predicting in-hospital mortality in the external validation population. However, both models had good calibration (greater than 0.80). Using routine neonatal data in a low resource setting it is therefore possible to predict in-hospital mortality using

either treatments or signs and symptoms. Using treatments as predictors (NETS) had the advantage of availability of data (0.8% data loss in complete case analysis) in contrast to the signs and symptoms (SENSS) which required imputation (19% data loss if complete case analysis was used). In conclusion, SENSS performs better than NETS in predicting in-hospital mortality but NETS is potentially more useful for routine practice in similar settings due to availability of predictors. Prediction of in-hospital mortality can be used to inform treatment decisions for individual patients and for case-mix adjustment as described in Chapter 1 (Table 1.9). I present a more detailed discussion of the clinical utility and the implications thereof for these models in Chapter 7 and 8.

# Chapter 7 : Evaluation of the utility of the NETS and SENSS models in clinical decision making

## 7.1 Introduction

### *7.1.1 Background*

In Chapter 5 and 6, I described the development and validation of the NETS and SENSS models to predict neonatal in-hospital mortality. The discriminatory ability of the final models after external validation is comparable: c-statistic 0.89 (95% C.I 0.86 to 0.92) and 0.89 (95% C.I 0.84 to 0.93) for NETS and SENSS respectively. The SENSS model is however better calibrated with a calibration slope 0.90 (95% C.I 0.78 to 1.01) compared to 0.76 (95% C.I 0.65 to 0.87) for the NETS model. Discrimination and calibration are however statistical measures of predictive accuracy and do not provide a direct estimate of clinical utility.(149) In this Chapter, I will evaluate and compare the clinical utility of the two models. Clinical utility is one of the factors that affect the adoption of prediction models into routine practice. I will therefore begin with an overview of factors that affect the adoption of prediction models by clinicians.

### *7.1.2 Factors affecting the adoption of clinical prediction models into routine practice*

The proliferation of prediction models as illustrated in Figure 1.7 of Chapter 1 has not been accompanied by a commensurate uptake in routine clinical practice.(93, 269) The clinical usefulness and uptake of a particular model is determined by factors that affect the four phases of model development, specifically: (1) development (2) validation (3) impact assessment and (4) implementation which

are shown in Figure 1.8 of Chapter 1.(269) Some of these determinants, derived from literature, are summarised in Table 7.1.

***Table 7.1**: Factors affecting clinical uptake of prediction models*

| Phase of development | Factors |
|---|---|
| **1.Development** | Evidence that model was developed using appropriate study design, statistical methods and adequate data.(269) |
| **2.Validation** | Evidence that model predictions are valid outside of the context in which it was developed (external validation).(269) |
| **3.Impact assessment** | Evidence of the effects (utility) of using model on decision making, patient outcomes and costs.(269, 299) |
| **4.Implementation** | Support by clinical experts in its field of application.(269) |
| | Predictors readily available in routine care.(269) |
| | Number of predictors in the model.(269) |
| | Format of the model e.g. regression formula, score chart, nomogram.(269, 299, 300) |
| | Ease of use at the point of care.(93, 136, 269, 300, 301) |
| | Saves time.(300, 301) |
| | Creates clinically meaningful groups that aid in decision making.(299-301) |
| | Relevance to the clinical context.(269, 302) |
| | Preference for own judgement over standardised practice.(93, 136, 269, 299, 302) |
| | Fear of medical legal consequences after decisions based on model.(136, 269, 299) |
| | Update of models as performance may wane over time.(269) |
| | Quality of reporting.(92, 269) |
| | Model is a standard of clinical care.(299, 301, 302) |
| | Face validity/clinically sensible.(93, 136, 299, 300) |
| | Lack of awareness.(302) |
| | Concern that important predictors are not included in the model.(136, 299) |
| | Perceived trade-off between efficiency and patient safety.(136, 299) |

In Chapter 5 and 6, I described in detail the development and validation of the NETS and SENSS models taking into consideration the factors that arise at these phases outlined in Table 7.1. Both models contain a small number of predictors (seven each) and a key consideration in their selection was availability in routine care thus addressing two factors under implementation in Table 7.1. In addition, as described in Chapters 3, 5 and 6, the models predict in-hospital mortality making

them relevant to the clinical context of neonatal units in LMICs where there is a high burden of neonatal mortality. In this Chapter, I focus on two further determinants of clinical uptake of prediction models that arise at the impact assessment and implementation phase: potential clinical utility and the format of model presentation respectively.

In the next section, I describe the estimation of potential clinical utility using decision curve analysis and the presentation of models using a simplified points system.(149, 150, 303, 304)

### *7.1.3 The role of decision curve analysis and presentation format in the clinical utility of prognostic models*

*a) Decision curve analysis*

In Chapter 2 (Section 2.7), I described in detail the computation and interpretation of decision analysis curves in the evaluation of the potential clinical utility of prognostic models. The principal component of decision curve analysis is the estimation of the "net benefit" of the models under consideration.(149, 150, 303) The net benefit quantifies the balance between benefits and harms across the range of predicted probabilities obtained from the models. This is computed using the formula shown in Box 2.5 of Chapter 2. Estimates obtained in this way provide data to : (1) compare the potential clinical utility of different models that predict the same outcome for a specified population (2) compare the model(s) against decision making without the model and (3) inform the creation of risk groups to guide decision making for individual patients based on the predicted probability of the outcome.(149) I describe in detail the evaluation of the clinical utility of NETS and SENSS using decision curve analysis in the methods section (Section 7.2).

*b) Model presentation*

There are various formats in which prediction models may be presented for clinical use. These include regression formulae, spreadsheets, nomograms, score charts, tables, and specific computer application programs.(136) There is however, a paucity of evidence on clinicians' preferences for model presentation.(136) The regression formulae for NETS (Box 5.3, Chapter 5) and SENSS (Box 6.2, Chapter 6) can be used to directly compute the probability of in-hospital death for individual patients based on their predictor profiles. If done manually, this is a tedious method given the decimal points of the predictor coefficients and the need to convert the linear predictor to probability of in-hospital death. Sullivan and colleagues have described a simplified points based system for the development of score charts.(304) This method simplifies the regression model allowing for the the probability of the outcome to be determined without a calculator or computer.(304) The steps are summarised in Box 7.1.

*Adapted from Sullivan LM, Massaro JM, D'Agostino RB. Presentation of multivariate data for clinical*

1. Estimate the parameters of the multivariable model
2. Organize the predictors into categories and determine reference values for each e.g. the mid-point:
    - This step only applies to predictors that are continuous variables
3. Determine the referent risk factor profile by choosing a base category for each risk factor:
    - This step only applies to risk factors that are continuous variables
    - Categorical and binary variables already have a base category in the model
4. Subtract the base category coefficient from the coefficient of each category for every predictor
5. Define a constant (fixed multiplier) for the points system i.e. the coefficient value that will correspond to one point
6. Determine points associated with each of the categories of the predictors
7. Determine the theoretical minimum and maximum points total (range)
    - Minimum: sum of the smallest values for each predictor
    - Maximum: sum of the highest values for each predictor
8. Determine risks associated with point totals for the range

***Box 7.1****: Development of a points system for a multivariable logistic regression model use: The Framingham Study risk score functions. Statistics in Medicine. 2004;23(10):1631-60*

I describe the application of these steps to derive simplified points for the NETS and SENSS models in the methods section (Section 7.2).

### *7.1.4 Rationale*

The final NETS and SENSS models demonstrated good predictive accuracy at external validation. Discrimination was similar but the SENSS model was better calibrated. Predictive accuracy is however not a measure of clinical usefulness which is a key determinant of whether prediction models are taken up into practice. I therefore used decision curve analysis to evaluate and compare the potential clinical utility of the two models. In addition, I applied the simplified points system by Sullivan et al to demonstrate how the two models could be used in neonatal units of low and middle-income countries to predict in-hospital mortality.(304)

### *7.1.5 Objective*

To evaluate the potential clinical utility of the NETS and SENSS models for predicting in-hospital neonatal mortality by:

a) Determining the net benefit for each model across the range of predicted probabilities using decision curve analysis
b) Comparing the net benefit of the two models against the default strategy of not using the models
c) Using a simplified points system to present the models for clinical use

## 7.2 Methods

In this section, I first describe the application of decision curve analysis to evaluate the potential clinical utility of the NETS and SENSS models in predicting in-hospital mortality.(149, 150, 303) As described in Chapter 1 (Section 1.6.3), prediction of in-hospital mortality may be applied in decision making at individual patient level, for groups of patients and service delivery planning and organization. In examining the clinical impact of NETS and SENSS, I have used the example of patient referral as it applies to individual patients and for service delivery planning. I thereafter describe the construction of score charts as a presentation format for the two models using the simplified points system by Sullivan et al.(304) For both the decision curves and score charts, I used the final NETS and SENSS regression models after updating at external validation described in Chapter 5 and 6 respectively.

### *7.2.1 Decision curve analysis*

I used the decision curve analysis approach to evaluate the potential clinical utility of the NETS and SENSS model in two ways. The first was calculation of the net benefit of each model with the corresponding decision analysis curves. The second was evaluation of the clinical impact of different cut-offs in classifying patients as high risk or low risk across the range of predicted probabilities.

#### *a) Computing the net benefit and plotting decision analysis curves*

The theoretical basis of decision curve analysis is described in detail in Chapter 2 (Section 2.7) and summarised in Section 7.1.3 (a) of this Chapter. Predicted probabilities of in-hospital death were obtained for the NETS and SENSS models

by applying their respective regression formulae on the external validation data set using the built-in function “predict.lm” of the R statistical programming language (version 3.3.3).(128) The results were two data sets containing observed in-hospital death and the corresponding predicted probability of in-hospital death for individual patients by the NETS and SENSS models. Using the “decision_curve” function of the "DecisionCurve" package of R, I then computed the net benefit across the range of predicted probabilities for each model.(151) Finally, I plotted the decision curves using the “plot_decision_curve” function of the "DecisionCurve" package in R.(128, 151) The code for these analyses is included in Appendix 9.

*b) Evaluating potential clinical impact*

The net benefit metric from decision curve analysis provides a measure of the balance between benefits and harms across the range of predicted probabilities for a given prediction model.(149) The curves thus visualise the range of predicted probabilities for a given model where benefit is greater than harm and thus clinically useful. For a model to guide decision making, we need to then select a cut-off for this range of predicted probabilities.(136) The clinical usefulness of the model is thus the ability to classify patients as “positive” (predictions above the cut-off) or “negative” (predictions below the cut-off).(136)

To examine potential cut-offs for the NETS and SENSS models I used the “plot_clinical_impact” function of the "DecisionCurve" package in R.(128, 151) The number of patients who would be classified as “positive” at each potential cut-off was plotted against the range of predictions for each model. In addition, the number with the observed outcome (in-hospital death) at each potential cut-off was plotted

on the same graph. This allows for the comparison of the number of predicted deaths ("positive") and observed deaths at each potential cut-off. The code for these analyses is included in Appendix 9. These curves provide information to guide selection of an optimal cut-off by a panel of clinicians for a particular context as cut-offs cannot be determined by statistical methods only.(136) This panel discussion is not a component of this D.Phil. but is considered further in Chapter 8.

*7.2.2 Simplified points based system*

I developed a score chart for the NETS and SENSS models as outlined in Box 7.1. I first selected the reference category for the respective predictors in the regression models as the base category. For birthweight and sex, normal weight (2.5-4kg) and female are the respective base categories. The treatment predictors included in NETS are binary with treatment "not prescribed" set as the base category (Table 5.2, Chapter 5). Similarly, the symptoms and signs included in SENSS are binary therefore the base category is documented absence of the sign or symptom (Table 6.2, Chapter 6).

The next steps are illustrated in Table 7.2. For *each* predictor in the NETS and SENSS model (represented by "p"), I subtracted the coefficient of the baseline category ($W_{REF}$) from the coefficient of the other categories (x-0).

*Table 7.2: Computation of simplified points from the model coefficients*

| Predictor | Categories | Baseline ($W_{ij}$) | $\beta_i^*$ | $\beta_i$ ($W_{ij}$ − $W_{iREF}$) | $\beta_i$ ($W_{ij}$ − $W_{iREF}$)/B† |
|---|---|---|---|---|---|
| $p_{ij}$ | 0 | 0=$W_{REF}$ | 0 | x-0 | (x-0)/B |
| | 1 | | x | | |

* $\beta_i$ regression coefficients
† Division by constant (B) and rounded off to nearest integer
Note: Adapted from Sullivan LM, Massaro JM, D'Agostino RB. Presentation of multivariate data for clinical use: The Framingham Study risk score functions. Statistics in Medicine. 2004;23(10):1631-60.

I then selected the smallest coefficient from each model as a constant (B) and divided the difference of the coefficients (x-0) by this constant. The quotient was then rounded off to the nearest integer to give points for the respective model coefficients. I summated the least points from each predictor to give the theoretical minimum number of total points from the model. Similarly, the sum of the highest points from each predictor was the maximum total points. This minimum and maximum defines the range of points for each model.

Every point on this range was then converted to a corresponding probability of in-hospital death in two steps as illustrated in Box 7.2. First the points for each model were multiplied by the constant (B) and each result was added to the intercept for the respective models: -4.877 for NETS (Table 5.13, Chapter 5) and -4.1946 for SENSS (Table 6.11, Chapter 6). The results were then converted to a corresponding probability using the formula in Box 7.2.

**1. p= (points*B) + model intercept**

**2.Probability of death = $1/(1+e^p)$**
Where $e^p$ =the natural log exponent of p

***Box 7.2**: Conversion of model points to predicted probability of in-hospital death*

The simplified points and the corresponding predicted probabilities were then presented as a score chart. In the next section, I present the results of the decision curve analysis and the points system with examples of their clinical application.

## 7.3 Results

### *7.3.1 Decision curve analysis evaluating clinical utility of the NETS and SENSS models*

*a) Comparison of the net benefit of the NETS and SENSS models*

Figure 7.1 shows the decision curves for NETS and SENSS models in predicting the risk of in-hospital death. The x-axis is the range of predicted probabilities of in-hospital death. This is labelled as “threshold probability” as each prediction in the range is considered a potential cut-off for classification of patients as high or low risk of in-hospital death.

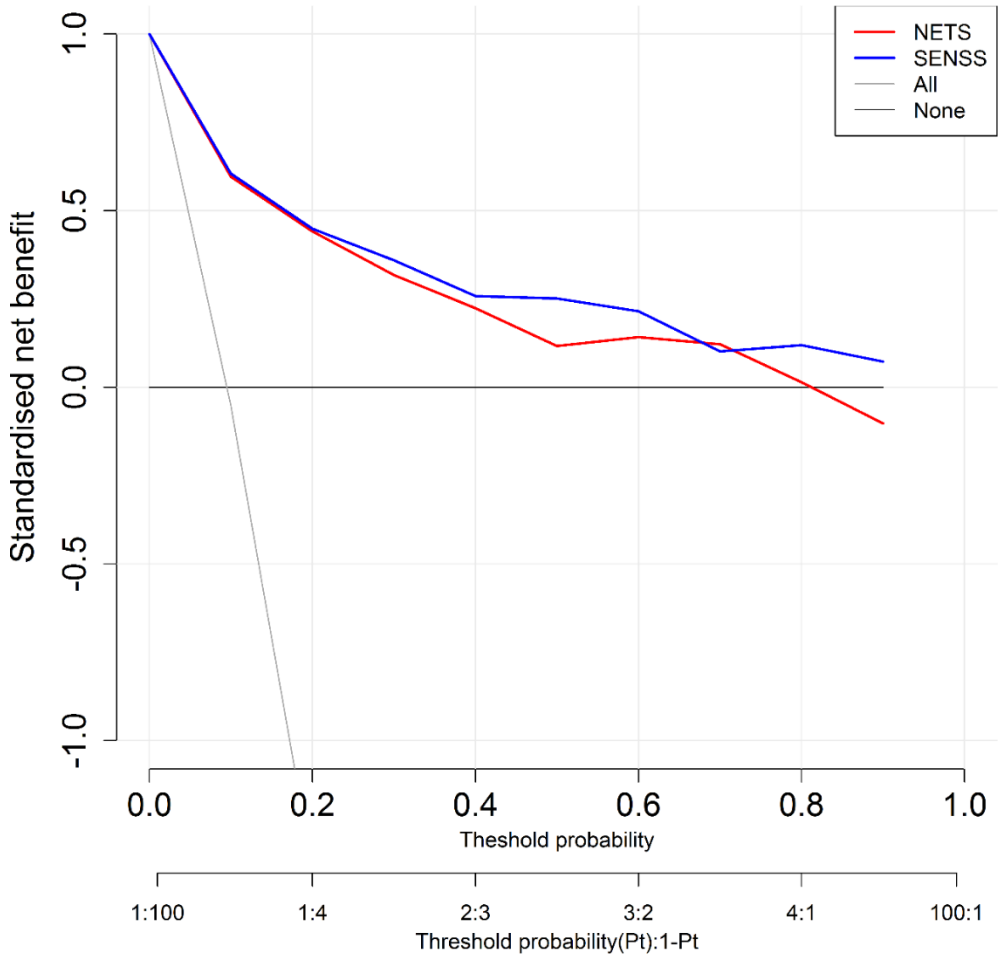



***Figure 7.1**: Comparison of the standardised net benefit treated for NETS and SENSS models in predicting in-hospital mortality*

Below the x-axis are the weights applied to the formula shown in Box 2.5 of Chapter 2 when calculating the net benefit at each threshold probability. This is the ratio of the threshold probability (Pt) to one minus the threshold probability (1-Pt).(149)

The y-axis shows the standardised net benefit (sNB) computed across the range of predicted probabilities of in-hospital mortality. This considers the perspective of patients who will receive an intervention based on the predicted risk of in-hospital death i.e sNB treated. The standardised net benefit (sNB) is obtained by dividing the net benefit (NB) at each threshold probability by the maximum net benefit of the model under evaluation.(305) The maximum NB is equal to the observed prevalence of the outcome of interest in the study population.(305) This is because for a perfect model, there are no false positives and the maximum proportion of true positives is equal to the observed population prevalence of the outcome. This means the sNB is on a scale of 0 to 1. A standardised net benefit greater than zero means benefit is greater than harm (or cost) and there is clinical utility. In the population depicted in Figure 7.1, the prevalence of in-hospital mortality was 9.3% therefore all the values of NB computed using the formula in Box 2.5 of Chapter 2 were divided by 0.093. Given that the maximum sNB is 1, each value of sNB may be interpreted as a proportion of the maximum sNB.

In Figure 7.1 for example, the threshold probability of 0.4 corresponds with a sNB of about 0.25 for the SENSS model. This means that using 0.4 as a cut-off for the SENSS model will achieve about a quarter of the maximum attainable NB.(305) In addition,as described in Chapter 2 (Section 2.7) this means that if the threshold probability for intervention based on predicted risk of death is set at 0.4 for the

SENSS model, it would be equivalent to intervening in 25 neonates per 100 at risk of death and in-hospital death would have been observed in all 25.(149, 305)

For a given intervention based on the risk of in-hospital death, the “All” line in Figure 7.1 shows the standardised net benefit of intervening in all patients without using either NETS or SENSS (i.e. assume all are at high risk of death). This explains why the line crosses the zero standardised net benefit line at 9% which is the crude case fatality ratio of the study population as described in Chapter 4. The “None” line in Figure 7.1 represents the assumption that none of the neonates is at risk of in-hospital death. Under this assumption, none of the patients will experience benefit or harm from an intervention based on the risk of in-hospital death therefore the standardised net benefit is zero.

The two curves for the NETS and SENSS models therefore show the standardised net benefit of intervention based on risk of in-hospital death across the range of predicted probabilities in comparison to intervening in all patients without using either model (“All”) and intervening in none (“None”). In addition, they demonstrate the difference in net benefit between the two models.

There are thus four potential strategies for decision making depicted in Figure 7.1: (1) intervene in all patients; “All” (2) intervene in none of the patients; “None” (3) use the NETS model and (4) use the SENSS model. According to Vickers Et al, “…the strategy with the highest net benefit at a particular threshold probability has the highest clinical value”.(149) We can therefore make the following observations for any given intervention based on risk of in-hospital death in this population of neonates:

i. The NETS model is clinically useful (positive standardised net benefit) for predicted probability of in-hospital death between 0 to 0.8
ii. The SENSS model is clinically useful across the whole range of predictions
iii. NETS and SENSS are superior to the ‘intervene for all’ strategy across the whole range of predictions
iv. NETS and SENSS are superior to the ‘intervene in none’ strategy across the whole range of predictions
v. SENSS is superior to NETS given its overall higher net benefit across the range of predictions

*b) Evaluating the clinical impact of the NETS and SENSS models*

In section 7.3.1 (a), I showed that SENSS is clinically useful across the whole range of predictions while NETS is useful in the range 0 to 0.8. In this section, I evaluate the clinical implications of different potential cut-offs using patient referral as the intervention as an example. Figure 7.2 shows the clinical impact of cut-offs across the range of predictions for NETS and SENSS respectively, for a theoretical population of 1000 patients.

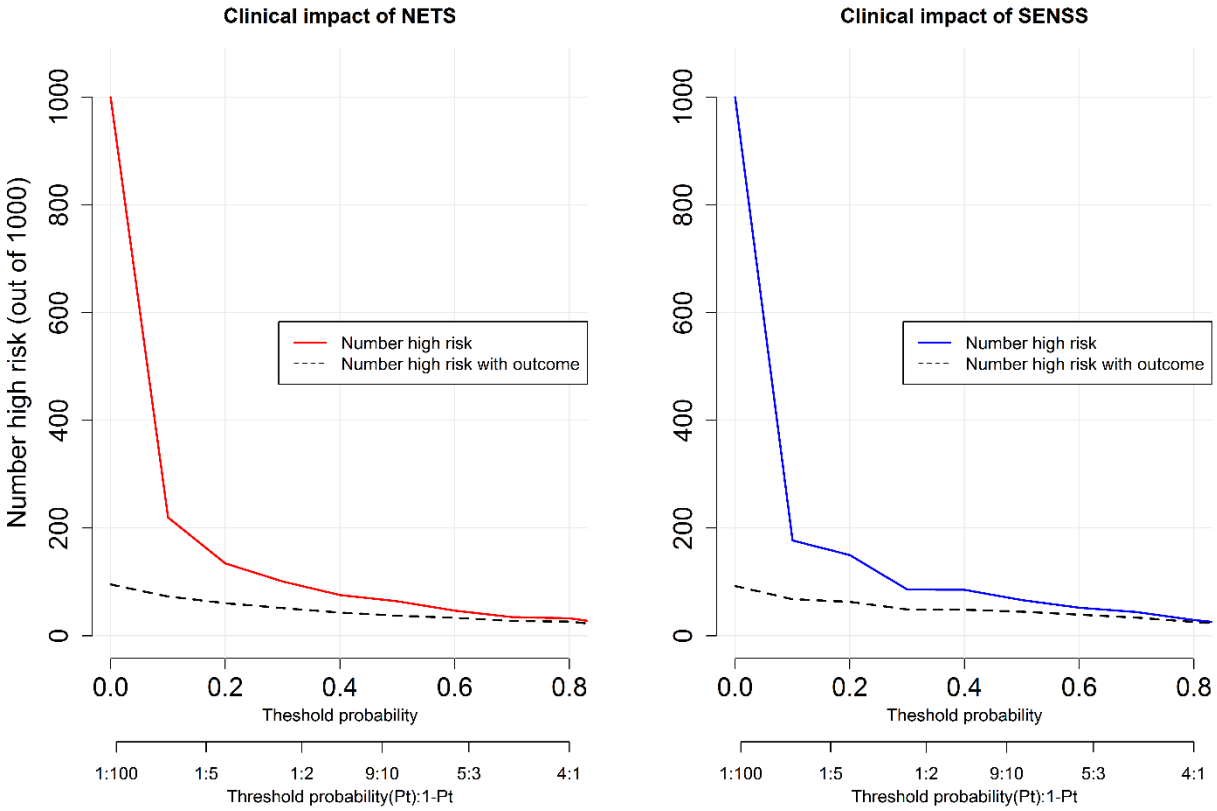



***Figure 7.2**: Clinical impact of cut-offs across the range of predictions for NETS (left) and SENSS (right)*

On the x-axis ("Threshold probability") is the range of predictions under consideration as potential cut-offs. For each cut-off as described in Section 7.2.1 (b) patients are classified as "positive" if they are at or above the cut-off or "negative" if below. The number "positive" at each cut-off are plotted on the y-axis

as “number high risk”. The solid line curves show the predicted number of in-hospital deaths at each cut-off across the range of predictions for the NETS (left side graph) and SENSS model (right side graph). The dotted line curves show the actual number of in-hospital deaths at each potential cut-off.

Figure 7.2 shows that as the cut-off increases, the number of patients at or above the threshold reduces for both NETS and SENSS. In addition, for both models as the cut-off increases the difference between the number at high-risk and the number of observed deaths reduces. For both models at the cut-off of 0.8, virtually all patients died in hospital. I shall now present as an example, the use of the NETS and SENSS models to determine which patients need referral based on the predicted risk of in-hospital death at admission.

There are three county hospitals in the public health system in Nairobi, Kenya one of which is the study site for this work. These hospitals refer sick neonates to the only tertiary level neonatal unit in the public health system in Nairobi. I will now consider the clinical implications of using the NETS and SENSS models to inform referrals from the three county hospitals based on the information in Figure 7.2. In 2015 these three facilities admitted approximately 5500 neonates. Table 7.3 illustrates the number of patients who would be referred based on the predicted risk of in-hospital death at different cut-offs.

***Table 7.3**: Estimated annual referrals at different cut-offs for NETS and SENSS in three Nairobi County hospitals*

| Cut-off* | NETS (n=5500) | | | | SENSS (n=5500) | | | |
|---|---|---|---|---|---|---|---|---|
| | Referred | (%) | Dead† | (%) | Referred | (%) | Dead† | (%) |
| **0.1** | 1210 | **22** | 400 | **33** | 990 | **18** | 387 | **39** |
| **0.2** | 770 | **14** | 347 | **45** | 825 | **15** | 347 | **42** |
| **0.4** | 440 | **8** | 251 | **57** | 495 | **9** | 282 | **57** |
| **0.6** | 275 | **5** | 198 | **72** | 275 | **5** | 206 | **75** |
| **0.8** | 165 | **3** | 134 | **81** | 165 | **3** | 146 | **88** |

* Cut-off for predicted probability of in-hospital death. All patients with predicted probability greater than or equal to the cut-off would be referred

† Observed number of deaths amongst those identified by each cut-off as referred. This is a subset of the number referred at each cut-off

As was illustrated in Figure 7.2, the higher the cut-off the smaller the number of patients who would be referred for both NETS and SENSS. Similarly, the proportion of potential referrals who die in-hospital increases with the cut-off. Table 7.3 illustrates the kind of data that would be presented to a panel of clinicians to inform the selection of optimal cut-offs for NETS and SENSS for a clinical context which in this case is referral decisions at the time of admission. Once a cut-off has been determined, the models need to be presented in a format that allows their use at the point of care. In the next section, I present the NETS and SENSS models as score charts for use by clinicians at the point of care.

*7.3.2 NETS and SENSS presentation as score charts*

In this section, I present score charts for the NETS and SENSS models. In addition, I illustrate the use of the score charts on real patients randomly drawn from the external validation data set.

*a) NETS model score chart*

Table 7.4 shows the simplified points system for NETS. For each predictor category, there are corresponding points reflective of its weight in the prediction model. The computations described in Section 7.2.2 resulted in the "Original points" in Table 7.4. I simplified these further by rounding off to the nearest ten, five or one to have numbers that would be easy to add up manually. These are labelled "Simplified points" in Table 7.4.

***Table 7.4**: Further simplified points system for the NETS model*

| Predictor | Categories | Original points | Simplified points |
|---|---|---|---|
| **1.Birthweight** | 2.5-4kg | 0 | 0 |
| | <1kg | 19 | 20 |
| | 1-<1.5kg | 15 | 15 |
| | 1.5-<2.5kg | 5 | 5 |
| | >4kg | -1 | -1 |
| **2.Sex** | Female | 0 | 0 |
| | Male | -1 | -1 |
| **3.Antibiotics** | No | 0 | 0 |
| | Yes | 5 | 5 |
| **4.Fluids** | No | 0 | 0 |
| | Yes | 5 | 5 |
| **5.Feeds** | No | 0 | 0 |
| | Yes | -7 | -5 |
| **6.Oxygen** | No | 0 | 0 |
| | Yes | 2 | 1 |
| **7.Phenobarbital** | No | 0 | 0 |
| | Yes | 9 | 10 |

For individual patients, the total number of points are determined by their predictor profile. This is illustrated in Section 7.3.2 (c). Table 7.5 shows the total points with

the corresponding risk of in-hospital death. The range for the original points is -9 to 40 and for the simplified -7 to 40.

***Table 7.5**: Comparison of predicted risk by the original and simplified points for NETS*

| Points | Risk Original (%) | Risk Simple (%) | Points | Risk Original (%) | Risk Simple (%) |
|---|---|---|---|---|---|
| **-9** | 0.1 | | **16** | 45 | 45 |
| **-8** | 0.1 | | **17** | 53 | 52 |
| **-7** | 0.1 | 0.1 | **18** | 60 | 60 |
| **-6** | 0.1 | 0.1 | **19** | 67 | 66 |
| **-5** | 0.2 | 0.2 | **20** | 73 | 73 |
| **-4** | 0.2 | 0.2 | **21** | 78 | 78 |
| **-3** | 0.3 | 0.3 | **22** | 83 | 83 |
| **-2** | 0.4 | 0.4 | **23** | 87 | 86 |
| **-1** | 0.6 | 0.6 | **24** | 90 | 90 |
| **0** | 0.8 | 0.8 | **25** | 92 | 92 |
| **1** | 1 | 1 | **26** | 94 | 94 |
| **2** | 1 | 1 | **27** | 95 | 95 |
| **3** | 2 | 2 | **28** | 97 | 97 |
| **4** | 2 | 2 | **29** | 97 | 97 |
| **5** | 3 | 3 | **30** | 98 | 98 |
| **6** | 4 | 4 | **31** | 99 | 99 |
| **7** | 6 | 6 | **32** | 99 | 99 |
| **8** | 7 | 7 | **33** | 99 | 99 |
| **9** | 10 | 10 | **34** | 99 | 99 |
| **10** | 13 | 12 | **35** | 99 | 99 |
| **11** | 16 | 16 | **36** | 99 | 99 |
| **12** | 20 | 20 | **37** | 99 | 99 |
| **13** | 26 | 26 | **38** | 99 | 99 |
| **14** | 31 | 31 | **39** | 99 | 99 |
| **15** | 38 | 38 | **40** | 99 | 99 |

Given the similarity of predicted risks by the original and simplified points, I adopted the simplified points as the final points system. I thereafter grouped together points corresponding to risks less than 1% and greater than 98% to produce the final score chart shown in table 7.6. This is accompanied by a form for recording predictor profile and calculation of total points for individual patients (Table 7.7).

***Table 7.6**: NETS final simplified score chart*

| Points | Risk (%) | Points | Risk Simple (%) |
|---|---|---|---|
| **<1** | <1 | **16** | 45 |
| **1** | 1 | **17** | 52 |
| **2** | 1 | **18** | 60 |
| **3** | 2 | **19** | 66 |
| **4** | 2 | **20** | 73 |
| **5** | 3 | **21** | 78 |
| **6** | 4 | **22** | 83 |
| **7** | 6 | **23** | 86 |
| **8** | 7 | **24** | 90 |
| **9** | 10 | **25** | 92 |
| **10** | 12 | **26** | 94 |
| **11** | 16 | **27** | 95 |
| **12** | 20 | **28** | 97 |
| **13** | 26 | **29** | 97 |
| **14** | 31 | **30** | 98 |
| **15** | 38 | **>30** | 99 |

***Table 7.7**: NETS patient score sheet*

| Predictor | Categories | Points | Patient score |
|---|---|---|---|
| **1.Birthweight** | 2.5-4kg | 0 | |
| | <1kg | 20 | |
| | 1-<1.5kg | 15 | |
| | 1.5-<2.5kg | 5 | |
| | >4kg | -1 | |
| **2.Sex** | Female | 0 | |
| | Male | -1 | |
| **3.Antibiotics** | No | 0 | |
| | Yes | 5 | |
| **4.Fluids** | No | 0 | |
| | Yes | 5 | |
| **5.Feeds** | No | 0 | |
| | Yes | -5 | |
| **6.Oxygen** | No | 0 | |
| | Yes | 1 | |
| **7.Phenobarbital** | No | 0 | |
| | Yes | 10 | |
| | | **Total points** | |
| | | **Risk** | |

*b) SENSS model score chart*

Table 7.8 and 7.9 show the final score chart and patient score sheet respectively for the SENSS model derived using the same method applied to the NETS model. The SENSS points range from -5 to 100 as shown in Table 7.8. Total points less than one correspond to risk of in-hospital death of 1% while 60 points and above denotes risk greater than 99%. Table 7.9 illustrates how the predictor profile for the SENSS model may be recorded for an individual patient allowing for calculation of a SENSS score. The risk of in-hospital death is then determined by reading off the calculated SENSS score against Table 7.8.

***Table 7.8**: SENSS final score chart*

| Points | Risk (%) | Points | Risk (%) | Points | Risk (%) | Points | Risk |
|---|---|---|---|---|---|---|---|
| **<1** | 1 | **18** | 14 | **26** | 33 | **43** | 82 |
| **1** | 2 | **19** | 16 | **27** | 36 | **44** | 84 |
| **2** | 2 | **20** | 18 | **28** | 39 | **45** | 86 |
| **3** | 2 | **21** | 20 | **29** | 42 | **46** | 88 |
| **4** | 3 | **22** | 22 | **30** | 45 | **47** | 89 |
| **5** | 3 | **23** | 25 | **21** | 20 | **48** | 90 |
| **6** | 3 | **24** | 27 | **31** | 49 | **49** | 91 |
| **7** | 4 | **25** | 30 | **32** | 52 | **50** | 92 |
| **8** | 4 | **26** | 33 | **33** | 55 | **51** | 93 |
| **9** | 5 | **27** | 36 | **34** | 59 | **52** | 94 |
| **10** | 5 | **28** | 39 | **35** | 62 | **53** | 95 |
| **11** | 6 | **29** | 42 | **36** | 65 | **54** | 95 |
| **12** | 7 | **30** | 45 | **37** | 68 | **55** | 96 |
| **13** | 8 | **21** | 20 | **38** | 71 | **56** | 96 |
| **14** | 9 | **22** | 22 | **39** | 73 | **57** | 97 |
| **15** | 10 | **23** | 25 | **40** | 76 | **58** | 97 |
| **16** | 11 | **24** | 27 | **41** | 78 | **59** | 98 |
| **17** | 13 | **25** | 30 | **42** | 80 | **≥60** | 99 |

***Table 7.9**: SENSS patient score sheet*

| Predictor | Categories | Points | Patient score |
|---|---|---|---|
| **1.Birthweight** | 2.5-4kg | 0 | |
| | <1kg | 45 | |
| | 1-<1.5kg | 30 | |
| | 1.5-<2.5kg | 10 | |
| | >4kg | -5 | |
| **2.Sex** | Female | 0 | |
| | Male | -1 | |
| **3.Difficulty feeding*** | No | 0 | |
| | Yes | 10 | |
| **4.Convulsions*** | No | 0 | |
| | Yes | 10 | |
| **5.Indrawing†** | No | 0 | |
| | Yes | 15 | |
| **6.Cyanosis†** | No | 0 | |
| | Yes | 10 | |
| **7.Floppy/unable to suck†** | No | 0 | |
| | Yes | 10 | |
| | | **Total score** | |
| | | **Risk** | |

* Symptom reported by caregiver upon interview by clinician

† Sign observed by clinician

In Section 7.3.2 (a) and (b), I have illustrated the translation of the NETS and SENSS models respectively into score charts that may be used by clinicians at the point of care to determine risk of in-hospital death for individual patients. I will now demonstrate the application of these score charts on real patient data.

*c) Application of the NETS and SENSS score charts on real patient data*

Table 7.10 shows the calculation of the risk of in-hospital death for six patients randomly selected from the external validation data set.

***Table 7.10:*** *Predicted risk of in-hospital death using the NETS and SENSS score charts for six cases randomly drawn from the validation dataset*

| | Patient | Common predictors | | NETS* predictors | | | | | SENSS† predictors | | | | | Score§ | Risk‖ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Weight‡ | Sex | T1 | T2 | T3 | T4 | T5 | S1 | S2 | S3 | S4 | S5 | | |
| **1** | **Dead** | 0.6kg | Female | Y | Y | N | Y | N | Y | N | N | Y | Y | | |
| | NETS | 20 | 0 | 5 | 5 | 0 | 1 | 0 | | | | | | 31 | **99%** |
| | SENSS | 45 | 0 | | | | | | 10 | 0 | 0 | 10 | 10 | 75 | **99%** |
| **2** | **Dead** | 1kg | Male | Y | Y | N | N | N | N | N | N | Y | Y | | |
| | NETS | 15 | -1 | 5 | 5 | 0 | 0 | 0 | | | | | | 24 | **90%** |
| | SENSS | 30 | -1 | | | | | | 0 | 0 | 0 | 15 | 10 | 54 | **95%** |
| **3** | **Dead** | 1.6kg | Female | Y | Y | N | N | N | Y | N | N | Y | Y | | |
| | NETS | 5 | 0 | 5 | 5 | 0 | 0 | 0 | | | | | | 15 | **38%** |
| | SENSS | 10 | 0 | | | | | | 10 | 0 | 0 | 10 | 10 | 40 | **76%** |
| **4** | **Alive** | 2.3kg | Male | Y | N | Y | N | N | N | N | N | N | N | | |
| | NETS | 5 | -1 | 5 | 0 | -5 | 0 | 0 | | | | | | 4 | **2%** |
| | SENSS | 10 | -1 | | | | | | 0 | 0 | 0 | 0 | 0 | 9 | **5%** |
| **5** | **Alive** | 3.3kg | Female | Y | Y | N | N | N | N | N | N | N | Y | | |
| | NETS | 0 | 0 | 5 | 5 | 0 | 0 | 0 | | | | | | 10 | **12%** |
| | SENSS | 0 | 0 | | | | | | 0 | 0 | 0 | 0 | 10 | 10 | **5%** |
| **6** | **Alive** | 5.5kg | Male | N | N | Y | N | N | N | N | N | N | N | | |
| | NETS | -1 | -1 | 0 | 0 | -5 | 0 | 0 | | | | | | -7 | **<1%** |
| | SENSS | -5 | -1 | | | | | | 0 | 0 | 0 | 0 | 0 | -6 | **1%** |

NETS treatment predictors: T1: antibiotics, T2: fluids, T3: enteral feeds, T4: oxygen, T5: phenobarbital
SENSS predictors: S1: difficulty feeding, S2: convulsions, S3: indrawing, S4: cyanosis, S5: floppy/unable to suck
Y: predictor present
N: predictor absent
*NETS: Neonatal Essential Treatment Score
†SENSS: Score for Essential Neonatal Symptoms and Signs
‡: birthweight
§: total score which is the sum of the points for NETS and SENSS predictors respectively
‖: risk of in-hospital death

For each of the six patients sampled, Table 7.10 shows the observed outcome alongside the respective predictor profile for NETS and SENSS models with the corresponding predicted risk of death. Birthweight and sex are predictors for both

the NETS and SENSS models but have different points due to the different coefficients in the respective final regression model (Box 5.3, Chapter 5 and Box 6.2, Chapter 6 respectively). The points allocated per predictor is based on its presence or absence in each patient using the score chart in Table 7.7 for NETS and Table 7.9 for SENSS. The total score for NETS and SENSS for each patient is the sum of the respective predictor points. The predicted risk of in-hospital death is obtained by checking the total points on the score chart in Table 7.6 for the NETS model and Table 7.8 for SENSS.

Three observations may be made from this example of the application of NETS and SENSS on real patient data. First, birthweight accounts for a greater proportion of the score as weight reduces. Secondly, predicted risks are the same at the extremes i.e. 99% and 1% but vary in between. Finally, patient 3 and 4 belong to the same weight category i.e. LBW (1.5 to 2.499 kg) but exhibit markedly different predicted risks; 38% versus 2% for NETS and 76% versus 5% for SENSS. This further reinforces my argument presented in Chapter 4 on the need to split this weight category into two 500g intervals.

## 7.4 Discussion

### *7.4.1 Summary of findings*

In this chapter, I have used a decision analytic approach and a simplified points system to evaluate the potential clinical usefulness of the NETS and SENSS prognostic models. Figure 7.1 shows that in clinical scenarios that consider risk of in-hospital death predicted at the time of admission, the SENSS model is slightly superior to the NETS model as it has an overall higher net benefit. In addition, if decisions on interventions are to be based on risk of in-hospital death, NETS and SENSS are superior to not using the model under the assumptions that either all or none of the patients are at risk of death. Net benefit is observed over a wide range of predicted probabilities or SENSS (0 to greater than 80%) in comparison to NETS (0 to ≈80%).

The clinical impact of prediction models may be evaluated by examining different risk thresholds as cut-offs within the clinically useful range of predicted risks. For NETS and SENSS a cut-off will distinguish between patients at high risk (at or above the threshold) and low risk of death (below the threshold). An example of a clinical scenario where such a risk threshold may be applied is in the referral of patients. In Figure 7.2 and Table 7.3, I consider the clinical impact of using different risk thresholds to refer neonates at the time of admission. These analyses demonstrate that the higher the threshold, the higher the proportion of deaths observed amongst those referred. The converse applies for the lower thresholds.

*7.4.2 Clinical utility of prediction models*

*a) The net benefit of NETS and SENSS for treatment (intervention)*

The decision analytic approach to the optimal risk threshold for a range of predicted risk involves evaluation of costs (loss or harms) associated with classification of patients by a given cut-off.(136) Cost may include financial costs and adverse medical outcomes e.g. reduced quality of life.(136) As we saw in Chapter 2 (Section 2.7), this typically requires additional information on these costs.(150) In the context of in-patient neonatal care in LMICs typified by the study hospital, additional information such as admission costs may not be readily available. Using the decision curve method, I quantified "costs" using the net benefit without any additional information as shown in Figure 7.1.

This evaluation of the clinical utility of NETS and SENSS models is a novel application of the decision curve analysis in neonatal medicine in Kenya. None of the models included in the systematic review of neonatal treatment scores in Chapter 3 reported net benefit. In addition, a search conducted in PubMed combining the search string used in the systematic review in Chapter 3 with the additional terms "decision", "curve" and "analysis" (Appendix 11) did not yield any results.(165) While this is not an exhaustive search, it suggests that decision curve analysis has not yet been applied to the evaluation of prognostic models in neonatal medicine beyond Kenya.

In Chapter 6 (Section 6.1.1), I cited work done by Collins and colleagues comparing the QRISK1 and QRISK2 models versus the NICE version of the Framingham risk score to explain the importance of comparing prediction models where there are

multiple prediction models for the same outcome.(284) In subsequent work, updated versions of the QRISK2 were compared to the NICE Framingham risk score using decision curve analysis.(306) The clinical utility (net benefit) of QRISK2-2008 and QRISK2-2011 was found to be superior to the NICE Framingham equation.(306) In addition, at the risk threshold of 0.2 (20%) used in clinical practice, the NICE Framingham equation had no clinical utility in identifying those who are at an increased risk of developing cardiovascular disease.(306) This affirms the role of decision curve analysis in the evaluation of novel prediction models like NETS and SENSS.

In Chapter 6 (Section 6.4.2) we saw that with regards to statistical measures of accuracy, SENSS and NETS have comparable discriminatory ability (c-statistic 0.89 and 0.89 respectively) while SENSS had better calibration (calibration slope 0.90) compared to NETS (calibration slope 0.76). There are however no statistical tests of significance for this observed difference in predictive accuracy. In addition, there is no established threshold for good calibration analogous to 0.8 for discrimination. In this regard, the net benefit approach is therefore superior when comparing performance of prognostic models as according to Vickers and colleagues, “the rank order is more important than the size of the difference”. This means that SENSS has a superior clinical utility regardless of the magnitude of the difference in net benefit when compared to NETS.(149)

*b) Net benefit in the untreated and overall net benefit*

In this Chapter, I have explored standardised net benefit for treatment. This is where intervention (treatment) is provided when the predicted probability of in-hospital mortality is higher than a given threshold (the threshold probability, Pt). The standardised net benefit in the treated is therefore the difference between the benefit of treating those with the outcome (true positives) against the cost (or harm) of treating those without the outcome (false positives) across the range of predicted risks.

Standardised net benefit for NETS and SENSS can also be considered for individuals deemed ineligible to receive the intervention because the predicted probability of in-hospital mortality is lower than the threshold probability for intervention. This is the sNB for the untreated.(307) Across the range of predicted probabilities of in-hospital mortality, there will be individuals with predicted probability less than the threshold who either don't experience the outcome (true negatives) or experience the outcome (false negatives).Not intervening in the true negatives is beneficial but failure to intervene in the false negatives results in harm.(307) To compute net benefit in the untreated, true positive and false positive in Box 2.5 are substituted with true negative and false negative respectively. The proportion false negative is multiplied by the reciprocal of the weight used in Box 2.5 as shown in Box 7.3 below

$$\textbf{Net benefit (untreated)} = \frac{\textbf{True negatives}}{\textbf{N}} - \frac{\textbf{False negatives}}{\textbf{N}} \times \frac{\mathbf{1 - Pt}}{\textbf{Pt}}$$

***Box 7.3****: Calculation of net benefit untreated across the range of threshold probabilities*

For a given model, the net benefit in the treated is therefore not equivalent to the net benefit untreated.(307) However, in this Chapter the objective was to compare the clinical utility of NETS and SENSS and according to Rousson and colleagues for any two models, the difference between net benefit treated is equivalent to the difference between net benefit untreated.(307) A perfect model will give the maximum net benefit which is equal to one minus the proportion with the outcome of interest in the study population since there are no false negatives. I therefore did not compute the net benefit in the untreated.

Given that the Chapter objective was adequately addressed by the net benefit treated, I also did not compute the overall net benefit. This is the sum of the net benefit for the treated and untreated across the range of plausible predicted probabilities.(307) For a perfect model, there are no false positives and false negatives and therefore the overall net benefit is at its maximum. The maximum overall net benefit is therefore interpreted as all individuals (treated and untreated) benefiting from decisions informed by the prediction model under evaluation with no harm or costs experienced.(307)

*c) The clinical impact of NETS and SENSS*

Based on the findings in Chapter 5 and Chapter 6, NETS and SENSS models perform well in predicting in-hospital neonatal mortality at the time of admission. In this Chapter, I have shown that the clinical utility of SENSS is superior to NETS based on net benefit. A clinician will however still be left with the challenge of translating this information into decision-making.(308) Making clinical decisions based on predicted risk of in-hospital death using either the NETS or the SENSS

models will require definition of a prediction score cut-off or threshold.(308) The clinical impact curves shown in Figure 7.2 shows the number of patients who would be classified at high risk of in-patient death across the range of predicted probabilities where clinical utility has been demonstrated using the net benefit.

A similar approach was used in the United Kingdom to examine how many of 1000 people would be identified as being at high risk at different risk thresholds using either QRISK2-2011 or the NICE Framingham equation, and how many of these go on to experience a cardiovascular event compared with a strategy where all individuals are deemed at high risk.(306) I applied this method to evaluate the potential use of NETS and SENSS to inform decisions on referrals at the time of admission to the neonatal unit in Nairobi County. The Kenyan national referral guidelines define a referral system as “A comprehensive health care system used to manage client health care needs by referring clients from an initiating facility to an organization, service, or community unit that can better provide the level of care needed”.(41)

Regarding referral of neonates in Nairobi, Table 7.3 shows how many neonates would be referred from three County hospitals to the tertiary National neonatal facility in Nairobi based on the risk of in-patient death at different risk thresholds by the NETS or SENSS models. Such data on the potential number of referrals can inform planning on bed capacity and other resources at both the initiating and receiving facility. This illustrates how prediction of risk of in-hospital death can inform clinical decisions for individual neonates (referral of those at high risk of death) and service planning (planning for the number of anticipated referrals). Such

decisions would however require the selection of a cut-off for NETS and SENSS. From a clinical perspective, at the extremes of predicted risk of death shown in Figure 7.2 and Table 7.3, NETS and SENSS are unlikely to be useful as risk is likely to be clinically apparent. This leaves a region of intermediate risk where there is likely to be more clinical uncertainty on risk of death where the prediction models would be useful. The optimal cut-off for referral would therefore lie within this range. Alternatively, rather than a single cutoff, this region of intermediate risk could define the range for optimal referral where low risk patients are cared for in lower level facilities while patients in this range are referred to higher level facilities. Such a strategy would strengthen the referral of patients for whom the higher level facility would have the highest impact.

*d) The role of decision curve analysis in selecting risk thresholds for NETS and SENSS*

In Section 2.7 of Chapter 2, I noted that decision curve analysis provides a simplified means to decision analysis since external information is not requ but does not determine the risk threshold for a given model. In this chapter, I have evaluated the standardised net benefit (treated) of NETS and SENSS across the range of predicted risks and considered the potential clinical impact in terms of numbers of patients classified as “high risk” using patient referral as an example. This information is however insufficient for selection of risk thresholds as these are dependent on the clinical context.(149) Selection of cut-offs is further discussed in Section 8.5.3 of Chapter 8.

*7.4.3 Clinical presentation of prediction models*

An important aspect of usability and therefore uptake of prediction models is their format of presentation.(269, 299, 300) In the Kenyan public health system, in-patient neonatal records are paper based just as in the study neonatal unit. The use of NETS or SENSS models in this context would therefore require a presentation format that integrates into a paper based workflow. Paper versions of the NETS and SENSS score charts (Table 7.7 and Table 7.9 respectively) could conceivably be filled in by clinicians at the point of admission. This has the potential advantage of face validity as the clinician can clearly see how each of the predictors is contributing to the predicted risk of in-hospital death. However, an extra form would mean extra work for the clinician which could negatively impact uptake. The advent of electronic medical records provides an opportunity to incorporate NETS or SENSS calculators to automatically compute risk of in-hospital death after the clinician completes the admission EMR thus avoiding the extra work load. (309)

*7.4.4 Limitations*

Decision curve analysis and assessment of clinical impact provides a range of predicted probabilities where the NETS and SENSS models are likely to be clinically useful. However, these analyses do not provide cut-offs for clinical application of the models. Instead they proffer measures that can be used to inform clinicians and other decision makers in selecting an optimal cut-off. In addition, these measures of clinical usefulness are not estimates of the effect of using the two models on patient outcomes. Such estimates can only be obtained in an impact study where there is prospective demonstration that use of the NETS or SENSS

model for decision making improves patient outcomes or system outcomes like appropriateness of referral.(299)

### *7.4.5 Implications*

NETS and SENSS are potentially clinically useful in predicting neonatal in-hospital mortality in hospitals providing essential neonatal care. In such facilities, the models may be presented as simple paper based score charts. These novel models are based on predictors available in basic routine data that does not include electronic physiological monitoring, laboratory or imaging evaluations. The NETS model suffers least from missingness that arises from poor medical record keeping.

## Chapter 8 : Discussion

### *8.1 Summary of findings*

Insufficient reduction of neonatal mortality in the MDG era (1990-2015) resulted in a growing relative contribution of neonatal mortality to under 5 mortality.(4) More than 98% of neonatal deaths occur in LMICs and hospitals in these regions are expected to play a key role toward achievement of SDG 3.2 “reduce neonatal mortality to 12 per 1000 live births and below”.(5, 310, 311) As described in Chapter 1, one approach in efforts to achieve this is to use data for decision making to improve quality of hospital care for neonates.

In this thesis, I have described clinical outcomes using routine data from a large urban neonatal unit in Nairobi, Kenya and explored the potential utility of simple prognostic models derived from these data in informing decision making for essential neonatal services. In the next two sections, I summarise the key findings as important evidence that may be applied to improve clinical outcomes in line with the Prognostic Research Strategy (PROGRESS) framework described in Chapter 1.(68) These results fall under three of the four PROGRESS themes: (1) fundamental prognosis, (2) prognostic factors and (3) prognostic modelling and are likely the first application of these recommendations in an African context.(68)

#### *8.1.1 Fundamental prognosis*

In Chapter 4, I observed that about 1 in 10 of the neonates admitted to the study facility died (839/9115, 9.2% [95% C.I 8.7 to 9.8%]) which is on the lower side of the range in the 15 reports from Kenya (ranging from 3 to 62%, Table 1.8) and 23 reports from hospitals of other African countries (ranging from 3 to 35%, Table 4.1)

identified in the literature. The overall median length of stay was 2 days with a range of 0 to 98 days (IQR 1 to 5days). The overall case fatality and median length of stay however masks considerable differences by weight category. The smallest neonates by birthweight (<1kg), experienced short stays (median 1 day) and the highest case fatality ratio (92%, 45/49) in contrast to the largest (≥ 4kg) with similar short stays (median 1 day) but the lowest case fatality (2%, 7/353).

The outcome for the ELBW(<1kg) is likely due to lack of intensive care in the study unit for example limited access to CPAP and neither surfactant nor mechanical ventilation are available contributing to rapid high mortality. The Neonates weighing more than 4kg are largely stable babies admitted for a short period of observation due to risks associated with this weight e.g. low blood glucose levels.(312) This contrasts with routine data from two neonatal networks in high income settings shown in Table 8.1

***Table 8.1**: Case fatality by birthweight in the study hospital versus two high income setting neonatal networks*

| Weight | Case Fatality Ratio | | | | | |
|---|---|---|---|---|---|---|
| | Study hospital | | ANZNN* | | CNN† | |
| | n | % | n | % | n | % |
| **<500** | 45/49 | 92 | 10/26 | 38 | 23/40 | 57 |
| **500-749** | | | 89/378 | 23 | 93/406 | 23 |
| **750-999** | | | 72/651 | 11 | 61/680 | 9 |
| **1000-1249** | 192/280 | 69 | 22/805 | 3 | 32/792 | 4 |
| **1250-1499** | | | 24/996 | 2 | 17/864 | 2 |
| **1500-1999** | 157/670 | 23 | 41/1610 | 2 | 46/4672 | 1 |
| **2000-2499** | 100/1448 | 7 | 45/1258 | 4 | | |
| **2500-2999** | 336/6253 | 5 | 43/1254 | 3 | 74/7359 | 1 |
| **3000-3499** | | | 45/1258 | 4 | | |
| **3500-3999** | | | 32/943 | 3 | | |
| **≥4000** | 7/353‡ | 2 | 9/493 | 2 | | |
| **Overall** | 839/9115 | 9 | 435/9672 | 4% | 370/14813 | 2.5 |

* ANZNN: Australia and New Zealand Neonatal Network (data for 2015 from 28 neonatal units)
† CNN: Canadian Neonatal Network (data for 2015 from 30 neonatal units)
Note: the shaded regions represent the five weight categories typically used in clinical practice :<1000g, 1000-1499g, 1500-2499g, 2500-4000g, >4000g

Table 8.1 compares case fatality ratio by birthweight category in the study hospital against the Canadian Neonatal Network (CNN) and the Australia and New Zealand Neonatal Network (AZNN) reports for 2015.(258, 259) In both AZNN and CNN, case fatality is much lower than in the study hospital. The most marked differences are in the low birth weight categories (<2500g) where there is a stark difference for the VLBW (1000 to 1499g) with 69% in the study hospital against less than 5% in both AZNN and CNN. Similar high case fatality for VLBW has been noted in previous work in 22 Kenyan hospitals (51%, 95%C.I 32 to 70%).(32) This is therefore a priority group as a lot of these deaths could potentially be averted through high quality essential neonatal interventions.

Beyond crude mortality ratio which is typically all that is reported in literature from LMICs as described in Chapter 4, competing risk survival analysis (in-hospital death versus discharge alive) revealed that as expected the probability of being

discharged alive increases with increasing birthweight category (Figure 4.10 to 4.15). I split the LBW (1.5-2.499kg) - which is the widest of the three categories below 2.5kg- into two 500g subcategories to investigate where the probability of discharge alive rises above probability of in-hospital death. Figure 4.14 shows that the switch occurs in the 2 to 2.499kg subcategory. This suggests that in research reports and routine statistics from LMICs hospitals we ought to have at least these two extra (1.5 to 1.99 and 2 to 2.499kg) or even smaller categories such as the 250g categories used in ANZNN and CNN.(258, 259)

*8.1.2. Prognostic factors and prognostic modelling*

In Chapter 3, I described a systematic review of neonatal treatment intensity scores showing that such scores show promise in predicting mortality and morbidity but there is low certainty in the evidence on their performance for low resource settings. The evidence did however provide the rationale to examine if they could be usefully developed further for low resource settings such as Kenya where treatment data may be easier to obtain compared to other parameters like laboratory measures of physiological status. Availability of treatment data depends on prescriptions by clinicians only while capacity for laboratory testing is often not available in low resource settings meaning data on biological markers of illness severity may either not be available at all or only available on a small (and likely biased) sub-group of newborns.

I therefore used essential treatments prescribed at admission as predictors to develop and validate a treatment intensity model to predict in-hospital mortality; the Neonatal Essential Treatment Score (NETS) described in Chapter 5. I compared

this model to a second model based on symptoms and signs of severe illness included in the national guidelines to aide treatment decisions; the Score for Essential Neonatal Symptoms and Signs (SENSS) described in Chapter 6. The set of clinical signs employed in the latter is based on the best available evidence on signs that are appropriate to the Kenyan context and that can inform admission decisions.(51) Both final models have good discrimination (each having a c-statistic of 0.89) but the SENSS model is better calibrated (calibration slope 0.90 compared to 0.76 for NETS). Using treatments as predictors (NETS) is advantageous if we assume that any treatment that was not prescribed was deemed unnecessary by the clinician rather than unavailable or missing as described in Chapter 5. The SENSS model however required multiple imputation due to missingness of symptoms and signs (19% of observations would have been dropped by complete case analysis), consistent with previous work in Kenyan hospitals.(32) However, in the study hospital efforts to improve documentation are underway through audit of quality of documentation and quarterly feedback on performance supported by the data collection system described in Chapter 2. In conclusion, the SENSS model has better predictive accuracy than the NETS model in predicting in-hospital mortality but NETS is potentially more useful for routine practice due to availability of data (essential treatments prescribed at admission). Usability in routine practice is however a proposition that will require testing in particularly with regards to clinicians' preference as this will be a key determinant of uptake.

In Chapter 7, I showed that for decisions where risk of in-hospital death at the time of admission is an important consideration, the SENSS model is slightly superior to the NETS as it has a higher overall net benefit. The net benefit described in

Chapter 2 and 7, is a measure of potential clinical utility based on the balance between benefit and harms across the range of predicted risks.(149) The NETS and SENSS models are both superior to not using either prognostic model as an aide to prognosis with the implicit assumption that either all or none of the neonates are at risk of death at the time of admission. Clinical usefulness as measured by the net benefit is observed over a wider range of predicted probabilities for SENSS (0 to greater than 80%) in comparison to NETS (0 to ≈80%) due to the superior predictive accuracy of the former.

In the next sections, I discuss the implications of my work for neonatal prognostic research and its application to essential neonatal care in LMICs.

### *8.2 Implications for neonatal routine data and hospital based prognosis research*

In this section, I consider the contribution of my work to hospital based neonatal research in Kenya and more widely for LMICs with regards to the use of routine data in general and specifically the field of prognostic research.

#### *8.2.1 The potential of neonatal routine data in LMICs*

In Chapter 1 (Table 1.8) and Chapter 4 (Table 4.1), I showed gaps in the quantity and quality of neonatal data from hospitals in Kenya and other African countries respectively. These data show that there is no standardised reporting of neonatal outcomes in Kenya and most of Africa. The available literature cannot therefore be considered a coherent body of work on neonates from hospitals in the region. As part of my research, I helped develop and implement a data collection system that assembled what is likely the largest neonatal data set from routine hospital care in

Kenya. The data set contains 11,850 observations from 23 months in the period April 2004 to July 2016 (Table 2.9, Chapter 2). Published reports from Kenyan hospitals (Table 1.8) show the largest data set hitherto included 8756 neonates collected over 19 years in one County (district) hospital in Kilifi, while from African hospitals (Table 4.1) the largest was from one tertiary hospital in Cameroon with 7824 neonates over ten years.(25, 239)

In terms of data quality, the data collection mechanism described in Section 2.3.3 of Chapter 2 is a novel implementation in a neonatal unit of a system used in a Kenyan network of paediatric units spanning 14 hospitals.(281) The principal elements of data quality assurance in this system are: (1) improving routine clinical documentation in collaboration with clinical teams using a standard admission proforma (NAR, Appendix 10) coupled to audit and feedback (2) a robust informatics framework using open source software for data collection and management described in detail in elsewhere.(283, 313)

The data and data collection system thus demonstrates the opportunity routine data presents in filling the gaps in the quantity and quality of neonatal data available from hospitals in LMICs. This is evidenced by its successful application for general paediatrics across 14 hospitals in Kenya.(281, 282, 288, 314) Scaling up neonatal data collection in this manner will be in line with the global agenda to improve neonatal data captured in the Every Newborn Action Plan (ENAP) strategic objective five; count every newborn.(315) However, a key aspect of data quality that cuts across LMICs and high income countries that is yet to be addressed is the standardisation of definitions and terms. In the next section, I

describe efforts to standardise the definition of neonatal outcomes that may inform further improvements in data systems in LMICs over time.

*8.2.2 Standardising neonatal outcomes in research*

Differences in outcome definition and measurement between studies preclude comparison and aggregation of study results.(316) The classification of causes of neonatal death illustrates this well. The WHO Global Reference List of 100 Core Health indicators defines neonatal mortality rate but does not include classification of causes of death.(219) Accurate classification of the cause of death is necessary to inform prevention efforts at the population level and for informing families.(317)

There has been a multiplicity of classification systems for perinatal deaths predominantly from high income countries but most deaths occur in LMICs.(317) Consequently, the WHO has developed the International Classification of Diseases – Perinatal Mortality (ICD-PM) which includes 11 categories for classifying neonatal deaths in the perinatal period.(317) The ICD-PM is the first global neonatal classification system but will require significant resources for widespread implementation. A network of neonatal units in Kenyan hospitals is an example of a platform that would scale up the use of the system in a wider neonatal data system.

I have been involved in two further international initiatives to standardise the definition and reporting of neonatal outcomes in general beyond classification of cause of death. These are (1) Strengthening the Reporting of Observational Studies in Epidemiology for Newborn Infection (STROBE-NI) and (2) Core

Outcomes in Neonatology (COIN). The STROBE-NI project which is complete and has been published, provides guidance for reporting observational neonatal infection studies including outcomes.(318) The COIN project which is still underway, aims to develop a core outcome set for neonates to facilitate standardized reporting in both clinical trials and observational studies.(316) A core outcome set is defined as “a collection of important outcomes identified through robust consensus methods by key stakeholders”.(316) A registry of core outcomes sets for different areas of medical practice is maintained by The Core Outcome Measures in Effectiveness Trials (COMET) initiative.(319)

Application of global outcome definitions such as the ICD-PM, STROBE-NI and COIN will further enhance the quality of routine neonatal data from hospitals in LMICs and globally by allowing aggregation of data across time and place for comparison and tracking progress. This will be essential in scaling up routine data collection with systems akin to that implemented in the study hospital to strengthen hospital information systems to provide data that fulfils national and international reporting requirements.

*8.2.3 Impact on neonatal prognostic research*

There is a gap between the potential and actual impact of prognosis research on health.(68) A major contributor to this is the relatively poor quality of available prognosis research in comparison to for instance, the rigour now applied to the conduct of (good) randomized controlled trials.(68) My work advances the quality of prognosis research by the novel application of the PROGRESS recommendations to neonatal medicine in general and for hospitals in LMIC

specifically. The latter is particularly significant as the burden of neonatal mortality is greatest in LMICs and hospitals are a key platform for the delivery of essential interventions.(320) The impact of prognosis research in improving neonatal health is therefore likely to be greatest in these settings. For these results to have an impact will however, require their integration in clinical decision making, healthcare policy, and discovering and evaluating new approaches to patient management.(68) I discuss the utility of the study results in these decision making domains in the next section.

### *8.3 Prognostic data for decision making to improve neonatal survival in Kenya*

In Chapter 1 (Table 1.5) I described three levels of decision making and the gaps in the current decision making processes in Kenyan Hospitals. The three levels are: (1) individual patients (2) groups of patients and (3) service delivery planning and organisation. I will now explore the application of the study results in addressing the gaps in decision making processes with the aim of improving neonatal health in Kenya.

#### *8.3.1 Clinical decision making for individual patients*

One of the key tasks clinicians face is quantifying the severity of illness and using this information to make decisions on treatment (Table 1.9, Chapter 1). NETS and SENSS quantify illness severity at the time of admission by using treatments and clinical symptoms and signs respectively to predict risk of in-hospital death. The NETS model uses treatments as predictors and therefore cannot be used to influence decisions for treatments that are included in the model.

In contrast to the current approach of using individual symptoms and signs as a measure of severity of illness (Table 1.5, Chapter 1), the two models use multiple variables to estimate absolute risk of death for individual patients. This predicted risk of mortality could potentially be used to determine the level of care e.g. as an admission criterion for neonatal intensive care. Levels of care refers to risk appropriate categories of care ranging from basic care for low risk neonates to complex intensive care for high risk neonates.(321) A hospital may have a neonatal unit offering one level only but in larger facilities the unit may have different levels. Currently there are no criteria for determining admission into different levels of neonatal care in Kenya. Recent policy in Kenya indicates that neonates weighing <2kg should be admitted principally for Kangaroo Mother Care.(322) However this recommendation was not made on the basis of established risks and does not specify the level of care.

To fill this gap as more advanced neonatal care is developed in Kenya, predicted risk of in-hospital mortality could be used in a manner analogous to the recommendation on weight (<1500g) and gestation (32 weeks) by the American Academy of Pediatrics (AAP) for NICU admission.(321) In determining level of care, the NETS and SENSS models have the potential to impact on survival for individual patients given the evidence that suggests neonatal survival is associated with level of care and patient volume. In a study of 48,237 VLBW infants born in hospitals in California, USA, lower levels of care and lower number of infants receiving care were associated with higher odds ratios for death ranging from 1.19 (95%C.I 1.04 to 1.37) to 2.72 (95%C.I 2.37 to 3.12).(323) However, when neonates are assigned to the appropriate level facility the care provided has to be of good

quality for survival to be improved. Measuring quality of neonatal care in LMICs is in its infancy, being largely limited to descriptive reports of available resources and documented practices.(32, 324) Recent work from Kenyan hospitals however suggests that quality of care measured using a validated index of guideline adherence (Paediatric Admission Quality of Care score – PAQC score) was associated with mortality. Higher PAQC scores were associated with reduction in odds of death ranging from 9 to 30% per unit increase in the PAQC score.(325) More sophisticated measures such as these may therefore in future complement prognostic work in helping to explain neonatal outcomes.

Another potential application of NETS and SENSS is to guide therapeutic decisions for specific disease conditions. An example is Respiratory Distress Syndrome in preterm neonates. In this condition, BCPAP is a non-invasive method of respiratory support shown to reduce the need for more expensive and invasive interventions like mechanical ventilation.(326) In the current Kenyan guidelines, the Apgar score and birthweight are used as individual criteria (in addition to the Silverman-Anderson score described in Chapter 3) to exclude the use of limited BCPAP resources in neonates with high likelihood of poor outcomes (Apgar <4, birthweight ≤1000g) thus preserving limited capacity to those most likely to benefit.(17) In a context with no mechanical ventilation, BCPAP is likely most useful for neonates with intermediate severity of illness. In this clinical scenario NETS and SENSS scores could conceivably be used to measure severity of illness instead of birthweight and the Apgar. Application of the models to such clinical questions could be addressed in future research.

For any score to be useful at individual patient level, it must be successfully adopted in routine practice. However, few models are adopted as illustrated by a systematic review of models in obstetrics which identified 263 models developed between 1976 and 2012 none of which were used in routine practice.(93) In Chapter 7, I described factors that influence the uptake of prediction models in routine clinical practice. Two characteristics of the NETS and SENSS models make them particularly suited for routine practice in Kenya and potentially other LMICs: (1) The use of predictors which clinicians are familiar with and can be made available in routine practice 2) simplicity of presentation that aligns with work flow e.g. can be completed as part of admission evaluation and treatment planning.

The result is two models with good predictive accuracy that do not require parameters like laboratory evaluation of kidney function, blood gas analysis etc. which are typically not available in the study neonatal unit and similar units in Kenya and other LMICs. Even where laboratory tests are available, clinicians would typically have to wait for results to complete the score. The small number of predictors (seven in each model) allowed for creation of simple score charts that fit on one side of an A4 paper which could be presented as a wall chart or included as a sheet in the patient file. The charts are thus potentially well suited for clinical use in paper based medical record systems at the time of admission. They could also be automated within an electronic medical record when these become available

The ubiquitous use of the Apgar score in neonatal medicine could perhaps be explained by similarity in characteristics to the NETS and SENSS models. The Apgar score also consists of a small number of parameters (five clinical signs, Table 3.1) which are familiar to clinicians and obtained by observation only without the need for electronic monitoring or laboratory evaluation.(327) As described in Chapter 3, a low (<5) Apgar score at five minutes is associated with neonatal and infant death.(160) The Apgar score does not however provide an estimate of risk of death for individual patients unlike NETS and SENSS.(327) An additional advantage of the NETS and SENSS scores for newborn admissions is that the predictors are documented at the point of intended use (at admission to the neonatal unit) unlike the Apgar which is documented in the delivery suite.

The use of NETS and SENSS in making decisions for individual patients must however be examined in impact studies to evaluate the effect on outcomes such as mortality and length of stay. Impact assessment is the third step in prognostic model development (Figure 1.8) ideally conducted in a cluster randomized controlled trial.(328) If clinical impact is demonstrated, the final step is translation into routine practice (step four of Figure 1.8). Implementation strategies to influence clinicians behavior may include inclusion in clinical practice guidelines and use of audit and feedback to reinforce use of the models.(328)

*8.3.2 Describing and comparing cohorts of patients admitted to neonatal units*

In Chapter 4, I described case fatality, length of stay and time to discharge or death. These results are a significant addition to what is known about the outcomes of neonates admitted to neonatal units in Kenya and Africa given the paucity of data shown in Table 1.8 and Table 4.1 respectively. The Kaplan-Meir survival analyses and competing risk analyses of time to in-hospital death or discharge are new data for Kenya and possibly African neonatal units. The demonstration that competing risk analysis gives a more accurate estimate of time to death in a high mortality context means future reports on time to in-hospital death from neonatal units where mortality is high such as Kenya and other LMICs should adopt this method.

An important application of the case fatality, length of stay and time to death data will be for tracking progress over time. Regular reporting (e.g. annually) of such data will allow for tracking the impact of hospital care on mortality of small and sick neonates over time. As noted in Chapter 1, hospital outcomes data such as these will become an increasingly important metric at population level and globally as more deliveries take place in facilities. An example of use of routine hospital data to study the impact at a population level is a study of NICU admission rates over a 6-year period in the USA.(329) Admissions increased over the period but the authors raised concern that increased admission of normal weight (≥2500g) infants to NICUs may mean less critically ill infants were being exposed to costly (and possibly harmful) intensive care.(329)

The NETS and SENSS models may be used for case-mix adjustment when evaluating several neonatal units in a region e.g. Nairobi County. A study of eight

NICUs in Australia (7,212 inborn neonates) used predicted mortality from a logistic regression model for case-mix adjustment.(330) Observed mortality varied from 5.3% to 10.4% but after adjustment for case-mix, severity of illness and number of admissions the mortality rates were similar as no hospital exceeded the ±99.8% control limits on funnel plots.(330) While the unit of observation of outcomes was individual patients (n=7212), the unit of analysis was hospitals as mortality rate was summarised at hospital level.(331) With the wide control limits it is hard to have an observation outside the limits from a group of only eight hospitals.

Case-mix adjustment has been used to study variation of mortality in children aged more than one month following admission to 13 county referral hospitals in Kenya.(254) All-cause crude and cluster adjusted mortality rate was highly variable across the 13 hospitals and this variation was largely attributable to differences in severity of illness at the time of admission.(254) While prediction modelling was not used in this study it shows in principle the use of case-mix adjustment in the Kenyan context to help explore variation in outcomes. The NETS and SENSS models may therefore be potentially useful for case-mix adjustment in studies of neonatal units in Kenya. However, such use will require cautious interpretation given the well recognised limitations of standardised hospital mortality rates.(221, 332, 333) Such data are perhaps best utilised to prompt neonatal units to review their practices for quality improvement (as part of peer to peer networks) rather than rating performance.(330)

Besides case-mix adjustment, the need for pragmatic clinical pragmatic trials to address locally relevant clinical questions on essential interventions in in-patient

neonatal care in LMICs has been identified.(334) In such clinical trials, patients may be stratified by risk of death predicted by the NETS or SENSS to balance baseline risk of in-patient death between groups ensuring fairer comparison of interventions. In addition, in trials with all-cause mortality as an outcome, the case fatality ratios by birthweight in Chapter 4 will provide useful estimates of baseline risk.

*8.3.3 Service delivery planning and organisation*

Service delivery is one of the bottlenecks to delivery of essential neonatal interventions in LMICs.(335) In Chapter 1 (Section 1.8), we saw that the use of information to guide decision making is an essential component of rational design of hospital services to deliver quality neonatal care. Table 1.5 shows that currently, service design in Kenya is largely informed by experience and opinion. In this thesis, I have developed prognostic tools and in this section I explore their use in strengthening health systems through evidence based policy and capacity planning.

In the context of hospital care for neonates, demand and capacity planning is an essential component of service design both within a hospital and for hospitals within a health system. Using the example of the study unit in the Nairobi County Health system I will now illustrate novel application of prognostic data to inform demand and capacity planning in Kenya.

At hospital level, the length of stay data in Chapter 4 suggests that with improved survival of ELBW and VLBW, it is expected that a larger number of neonates will

stay in the study unit for a longer time. This has important implications on the capacity of the index unit to handle the anticipated demand. This is likely to impact on decisions to refer patients thereby impacting on the capacity of the Nairobi county health system also. The County health system should therefore be preparing for an increase in demand for neonatal care. In Chapter 7, I explored how the NETS and SENSS models could inform referral policy by providing rational referral criteria (based on risk of mortality) and the anticipated numbers of referrals that might result.

This is a unique illustration of how evidence from clinical epidemiology (predictive performance of NETS and SENSS) can feed into decision analysis for a health system policy, in this case referral policy and long term service planning. It is therefore one way that a systematic approach based on prognostic scores can inform efforts to improve the health of individual neonates and the population.(336) While I focused on up-referral of patients to the tertiary neonatal unit, the NETS and SENSS models could also be used at the tertiary facility to refer down patients at low risk of death to lower level neonatal units in Nairobi.

There is a real need for such data driven approaches to demand and capacity planning in Nairobi county and Kenya.(322) Given the labour intense nature of neonatal care, especially nursing care, such planning must include considerations for nursing (and medical) staffing. Population level estimates show that almost one in every five (18.3%, lower and upper level of confidence 14.8% to 22.1% ) liveborn babies in Nairobi County may require admission to a neonatal unit.(322) This translates to an estimated 24,161 (range 19,672 to 27,329) neonates in 2017.(322)

Unplanned expansion of capacity e.g. neonatal intensive care could lead to under use of expensive capacity rather than appropriate expansion of intermediate level facilities. A study of neonatal intensive care admissions across the USA found persistent underuse of NICU care for VLBW infants.(337) However, such planning has to go beyond the usual focus on bed capacity.(338) Bed number is not a good measure of services provided as determined by variation in case mix.(338) An alternative approach is the use of care pathways to understand and improve the flow of the patient through the whole healthcare system.(338) Prognostic data by providing information on anticipated demand e.g. length of stay and number of potential referrals can thus play a key role in demand and capacity planning for neonatal care in health systems of LMICs as part of care pathway analysis.

### *8.4 Strengths and limitations*

This work is an example of how research may be embedded within routine care to answer clinically relevant questions while strengthening the quality of routine data. These data are generated by multiple clinicians during routine practice over a period of two years rather than by researchers, enhancing the generalisability of the results and lowering the barrier to scale up in similar settings. The data set was large (>11,000) observations affording adequate events per variable (>10) and events versus non-events (>100 in each case) for prognostic model development and external validation respectively. In addition, the results obtained are a novel application of prognosis research methods thus advancing the state of the art for prognosis research in general and specifically for hospital based neonatal care LMICs.

In terms of limitations, this study was conducted in one facility potentially limiting generalisability of the results. However, the study hospital provides a significant proportion of hospital care for neonates in Nairobi thus making the results relevant to the County with the highest neonatal mortality in Kenya. Whereas NETS predictors are more likely to be available from routine data, we must assume that the prescriptions represent the admitting clinicians' full intent. However, the model is based on essential in-patient neonatal care which has a limited range of therapeutic interventions. This in turn potentially limits usability of NETS when more advanced interventions e.g. BCPAP and mechanical ventilation are introduced. However, this can be addressed by including the new interventions in the model and examining whether performance is improved.(339) Finally in the study data set there was no way of ascertaining if the treatments prescribed were actually given.

A key limitation of the SENSS model was the missingness of predictors. This was however remedied by conducting multiple imputation as recommended rather than a complete case analysis which can introduce bias and results in data loss.(101)

*8.5 Recommendations and future directions*

*8.5.1 Fundamental prognosis*

The collection and use of routine data to describe hospital outcomes for neonates should be expanded to other hospitals in Kenya and LMICs. This will go a long way in filling the gaps in data quantity and quality as we know very little about neonatal outcomes in Kenya. It will also allow the tracking of progress and facilitate the understanding of variation between and within countries aided by use of standard definitions. In this study I focused on mortality and length of stay. The results show that improved survival for ELBW and VLBW neonates will lead to longer hospital stays with attendant increase in costs (as they are likely to require high intensity care initially) as exemplified in Chapter 4 where 50% improved survival of ELBW will lead to 23 more neonates with a median length of stay of 23 days (from 49 with median stay of 1 day) and therefore more than 400 extra bed days of high dependence care. Secondly, competing risk analysis provides a better estimate of cumulative incidence of in-hospital mortality compared to Kaplan-Meier analysis and the clinical community should be informed on this so it becomes the standard in reporting. Thirdly, there is residual variation in risk of death in the LBW (1.5 to 2.499kg) category suggesting the need for better methods of estimating risk of in-hospital mortality such as prognostic modelling.

Beyond mortality, small and sick neonates who don't die are at risk of significant morbidity making non-fatal outcomes an important aspect of fundamental prognosis in neonatal medicine. The fundamental message in the series "Beyond Survival" published in Pediatric Research in 2013 is improved survival without a concomitant assurance of quality care will increase the risk of survival with

disability a phenomenon well recognized in high income countries.(340, 341). Non-fatal adverse outcomes e.g. neurodevelopmental impairment will thus become increasingly important for hospital based neonatal care in LMICs and future efforts to collect neonatal data should include these.(342-344)

### *8.5.2 Prognostic factor research*

Whilst I did not conduct primary prognostic factor research, I singled out the missingness of gestational age data as a key gap. Efforts to improve the accurate dating of pregnancy are therefore required in Kenya as gestation is a well known prognostic factor. This will also allow for accurate assessment of intrauterine growth. In the study neonatal unit, the routine reports provided as part of audit and feedback demonstrate improvement in documentation of gestational age. However, this is based on last menstrual period rather than the gold standard -first trimester ultrasound.(345) Routine first trimester ultrasound will however require whole system changes with major cost implications and will therefore take time to implement.

Small for gestational age (SGA) neonates are at increased risk of morbidity and mortality.(346) An estimated one in five infants are born SGA in LMICs (23 million in 2012, 19.3% of live births).(346) About a fifth (21.9%) of all neonatal deaths in these countries in 2012 was attributable to SGA.(346) Lack of gestational age data in LMICs results in failure to recognise this critical prognostic factor in areas where the burden is the largest. Gestation and assessment of appropriateness of size at birth are therefore priority areas for neonatal prognostic factor research especially in its application to case mix-adjustment.

Neonatal health is inextricably linked to maternal health as captured in the reproductive, maternal, newborn and child health (RMNCH) continuum of care concept.(347) While this concept is focused on integrating the delivery of essential interventions, it also provides a framework for important prognostic factors for neonates through the lifecycle and the health system with examples shown in Table 8.1.

***Table 8.2:*** *Examples of neonatal prognostic factors in the RMNCH continuum of care*

| Pre-pregnancy | Pregnancy | Intrapartum* | Postnatal |
|---|---|---|---|
| Nutritional deficiencies | Nutritional deficiencies | Preterm prelabour rupture of membranes | Initiation of breastfeeding |
| Maternal chronic illnesses | Maternal chronic illnesses | Skilled birth attendance | Newborn thermal care |
| Maternal infections | Maternal infections | | Newborn cord care |

*** During delivery**

Note: Adapted from Lassi ZS, Majeed A, Rashid S, Yakoob MY, Bhutta ZA. The interconnections between maternal and newborn health – evidence and implications for policy. Journal of Maternal-Fetal and Neonatal Medicine. 2013;26(S1):3-53.

Table 8.2 demonstrates that while focus on neonatal units in LMIC hospitals is vital, the spectrum of prognostic factors that affect neonatal outcomes goes beyond the neonatal unit. Collaborative efforts will therefore be required with reproductive health to build a complete profile of prognostic factors that extend to outcomes in infancy and childhood given the interconnectedness of RMNCH continuum.

### *8.5.3 Prognostic modelling*

NETS and SENSS are the most contextually appropriate prognostic models for Kenyan neonatal units offering essential in-patient care. The models have good performance in the study hospital after temporal external validation and can proceed to evaluation of impact for this unit. In other neonatal units however,

geographical external validation will be required. This will require data from other neonatal units in Nairobi and across Kenya to provide data sets for external validation. Recent work in 12 County hospitals in Kenya showed many neonatal admissions over a two-year period (November 2014 to November 2016). There were 24,212 admissions to the newborn units with case fatality (12%).(348) These are therefore potential sites for external validation. If the models show good performance after such geographical external validation a network of these neonatal units will provide an ideal platform to conduct a cluster randomized control trial to assess the impact of using the models on clinical outcomes.

As described in Chapter 7, cut-offs for different clinical scenarios e.g. in informing referral guidelines will have to be determined for the two Models. This will preferably be done by a panel of clinicians considering the decision analysis results in Chapter 7. An example recommendation would be to refer at admission all neonates with a NETS predicted risk ≥ 0.2 and < 0.4 to avoid referring patients at low risk of death (<0.2) and those at very high risk and unlikely to benefit from available interventions (≥0.4). There is however no consensus on how to define risk categories to be used for clinical decision making.(349, 350) Systematic reviews of prediction models show that investigators often do not provide a description of how risk categories were arrived at.(98, 349) In some cases, the cut-offs may actually be arbitrary.(349) Further, like categorisation of continuous variables, risk groups assume homogenous risk within categories.(349) In creation of risk groups therefore the following must be considered: (1) Cut-offs cannot be selected purely on statistical grounds, clinical context must be considered, (2) The

method used to arrive at the cut-offs should be reported, (3) Sufficient information should be provided to calculate individual risk.(149, 349)

The uptake of new technologies in neonatal units of Kenyan hospitals such as pulse oximetry(351) and BCPAP(251) provides opportunity to further develop locally applicable prognostic models in two ways: (1) inclusion of new predictors in NETS and SENSS, and (2) external validation of other neonatal models such as those described in Chapter 3 designed for NICU settings. Head to head comparisons of these models in local populations will then provide a means to selecting the best performing model for general application.(148)

Another issue going forwards for NETs and SENSS models is updating. Updating may be required in two contexts: (1) for new populations where deterioration in performance necessitates revision of the models and (2) where model performance deteriorates over time a phenomenon referred to as calibration drift.(352) The first case is described in Chapter 5 and Chapter 6 where I updated the model intercept for the NETS and SENSS model respectively as part of temporal external validation. In the second case, the causes of deterioration over time may include changes in disease prevalence, data quality, treatments and distribution of predictors and outcomes.(352) An example is the QRISK model for the prediction of 10 year risk of cardiovascular disease first published in 2007, updated in 2008 (QRISK2) and annually thereafter until 2017 when the QRISK3 was published.(353) Similar consideration will have to be made for NETS and SENSS over time if adopted into routine practice.

*8.6 Conclusion*

Using what is likely the largest routine neonatal dataset in Kenya, I have reported the first 'time to in-hospital mortality' analyses for Kenya and most likely the first competing risk analyses for neonates in African hospitals. This is significant as the competing risk approach is a more accurate measure of cumulative incidence of in-hospital mortality than the Kaplan-Meier method in a high mortality setting. The NETS and SENSS are novel prediction models that use relevant predictors, have good predictive accuracy and were developed only after a review of existing neonatal models revealed none was applicable to essential neonatal care. In addition, this work includes what is likely the first decision curve analysis of neonatal prognostic models globally. Decision curve analysis translates statistical performance measures to a clinically interpretable measure of utility (net benefit) which is likely to play a key role in facilitating adoption of NETS and SENSS into clinical practice. However, before introduction into practice the two models should undergo geographical external validation and impact assessment in a cluster randomised trial.

Prognostic research has an important role to play in the understanding and improvement of neonatal outcomes for hospital based in-patient neonatal care in LMICs. Good quality routine data collected across hospitals within and between countries can in this way contribute significantly towards the achievement of Sustainable Development Goal 3.2 "By 2030, end preventable deaths of newborns and children under 5 years of age, with all countries aiming to reduce neonatal mortality to at least as low as 12 per 1000 live births and under-5 mortality to at least as low as 25 per 1000 live births "and by extension Universal Health Care.

## References


1. Oza S, Cousens SN, Lawn JE. Estimation of daily risk of neonatal death, including the day of birth, in 186 countries in 2013: a vital-registration and modelling-based study. The Lancet Global Health. 2014;2(11):e635-e44.
2. Wang H, Bhutta ZA, Coates MM, Coggeshall M, Dandona L, Diallo K, et al. Global, regional, national, and selected subnational levels of stillbirths, neonatal, infant, and under-5 mortality, 1980–2015: a systematic analysis for the Global Burden of Disease Study 2015. The Lancet. 2016;388(10053):1725-74.
3. United Nations. The Millenium Development Goals Report. 2015.
4. Liu L, Oza S, Hogan D, Chu Y, Perin J, Zhu J, et al. Global, regional, and national causes of under-5 mortality in 2000-15: an updated systematic analysis with implications for the Sustainable Development Goals. The Lancet. 2016;388(10063):3027-35.
5. United Nations;General Assembly. Transforming our world: the 2030 Agenda for Sustainable Development 2015.
6. Lawn JE, Blencowe H, Oza S, You D, Lee ACC, Waiswa P, et al. Progress, priorities, and potential beyond survival. The Lancet. 2014.
7. United Nations DoEaSA. Word Population Prospects: The 2015 Revision,Volume 1:Comprehensive Tables.ST/ESA/SER.A/379. 2015.
8. World Bank. Gross domestic product 2015. 2017 28 April
9. United Nations Children's Fund. Number of neonatal deaths 2015 [updated January 2016. Available from: https://data.unicef.org/topic/child-survival/neonatal-mortality/.
10. Kenya National Bureau of Statistics, Ministry of Health Kenya, National AIDS Control Council Kenya, Kenya Medical Research Institute , National Council for Population and Development Kenya, The DHS Program ICF International USA Kenya Demographic and Health Survey 2014 2015.
11. Amouzou A, Kozuki N, Gwatkin DR. Where is the gap?: the contribution of disparities within developing countries to global inequalities in under-five mortality. BMC Public Health. 2014;14(1):216.
12. Indira N, Mandy R, Dilberth C, Silvana F, Tina S. The Components of Essential Newborn Care. 2004.
13. Kerber KJ, de Graft-Johnson JE, Bhutta ZA, Okong P, Starrs A, Lawn JE. Continuum of care for maternal, newborn, and child health: from slogan to service delivery. The Lancet. 2007;370(9595):1358-69.
14. World Health Organization Guidelines on Maternal,Newborn,Child and Adolescent Health: Recommendations on newborn health. 2013.
15. Bhutta ZA, Das JK, Bahl R, Lawn JE, Salam RA, Paul VK, et al. Can available interventions end preventable deaths in mothers, newborn babies, and stillbirths, and at what cost? The Lancet. 2014;384(9940):347-70.
16. Salam R, Mansoor T, Mallick D, Lassi Z, Das J, Bhutta Z. Essential childbirth and postnatal interventions for improved maternal and neonatal health. Reproductive Health. 2014;11(Suppl 1):S3.
17. Ministry of Health, Government of Kenya. Basic Paediatric Protocols for ages up to 5 years. 2016.
18. Opiyo N, Shepperd S, Musila N, English M, Fretheim A. The "Child Health Evidence Week" and GRADE grid may aid transparency in the deliberative process of guideline development. J Clin Epidemiol.65:962 - 9.
19. Barasa E, Ayieko P, Cleary S, English M. A multifaceted intervention to improve the quality of care of children in district hospitals in kenya: a cost-effectiveness analysis. PLoS Med.9(6):e1001238.

20. Milner KM, Duke T, Bucens I. Reducing newborn mortality in the Asia–Pacific region: Quality hospital services and community-based care. Journal of Paediatrics and Child Health. 2013;49(7):511-8.
21. Neogi SB, Malhotra S, Zodpey S, Mohan P. Does Facility-Based Newborn Care Improve Neonatal Outcomes? A Review of Evidence. Indian Pediatrics. 2012(49):651-8.
22. Woods DL. Improving neonatal care in district and community health facilities in South Africa. Paediatrics and International Child Health. 2015;35(3):187-91.
23. Tura G, Fantahun M, Worku A. The effect of health facility delivery on neonatal mortality: systematic review and meta-analysis. BMC Pregnancy and Childbirth. 2013;13(1):18.
24. Talbert AW, Mwaniki M, Mwarumba S, Newton CR, Berkley JA. Invasive bacterial infections in neonates and young infants born outside hospital admitted to a rural hospital in Kenya. Pediatr Infect Dis J. 2010;29(10):945-9.
25. Mwaniki MK, Gatakaa HW, Mturi FN, Chesaro CR, Chuma JM, Peshu NM, et al. An increase in the burden of neonatal admissions to a rural district hospital in Kenya over 19 years. BMC Public Health. 2010;10(591):1471-2458.
26. Neogi SB, Khanna R, Chauhan M, Sharma J, Gupta G, Srivastava R, et al. Inpatient care of small and sick newborns in healthcare facilities. J Perinatol. 2016;36(s3):S18-S23.
27. Dickson KE, Simen-Kapeu A, Kinney MV, Huicho L, Vesel L, Lackritz E, et al. Every Newborn: health-systems bottlenecks and strategies to accelerate scale-up in countries. The Lancet. 2014;384(9941):438-54.
28. Moxon S, Lawn J, Dickson K, Simen-Kapeu A, Gupta G, Deorari A, et al. Inpatient care of small and sick newborns: a multi-country analysis of health system bottlenecks and potential solutions. BMC Pregnancy and Childbirth. 2015;15(Suppl 2):S7.
29. de Graft-Johnson J, Vesel L, Rosen HE, Rawlins B, Abwao S, Mazia G, et al. Cross-sectional observational assessment of quality of newborn care immediately after birth in health facilities across six sub-Saharan African countries. BMJ Open. 2017;7(3).
30. Opondo C, Ntoburi S, Wagai J, Wafula J, Wasunna A, Were F, et al. Are hospitals prepared to support newborn survival? – an evaluation of eight first-referral level hospitals in Kenya. Tropical Medicine & International Health. 2009;14(10):1165-72.
31. Gathara D, Opiyo N, Wagai J, Ntoburi S, Ayieko P, Opondo C, et al. Quality of hospital care for sick newborns and severely malnourished children in Kenya: A two-year descriptive study in 8 hospitals. BMC Health Services Research. 2011;11(1):307.
32. Aluvaala J, Nyamai R, Were F, Wasunna A, Kosgei R, Karumbi J, et al. Assessment of neonatal care in clinical training facilities in Kenya. Archives of Disease in Childhood. 2015;100(1):42-7.
33. English M, Gathara D, Mwinga S, Ayieko P, Opondo C, Aluvaala J, et al. Adoption of recommended practices and basic technologies in a low-income setting. Archives of Disease in Childhood. 2014;99(5):452-6.
34. Warren C, Mwangi A, Oweya E, Kamunya R, Koskei N. Safeguarding maternal and newborn health: improving the quality of postnatal care in Kenya. International Journal for Quality in Health Care. 2010;22(1):24-30.
35. Diamond-Smith N, Sudhinaraset M, Montagu D. Clinical and perceived quality of care for maternal, neonatal and antenatal care in Kenya and Namibia: the service provision assessment. Reproductive Health. 2016;13(1):92.
36. Obiero CW, Seale AC, Berkley JA. Empiric treatment of neonatal sepsis in developing countries. The Pediatric Infectious Disease Journal. 9000;Publish Ahead of Print.
37. Musiime GM, Seale AC, Moxon SG, Lawn JE. Risk of gentamicin toxicity in neonates treated for possible severe bacterial infection in low- and middle-income countries: Systematic Review. Tropical Medicine & International Health. 2015;20(12):1593-606.

38. Crockett M, Avery L, Blanchard J. Program science—a framework for improving global maternal, newborn, and child health. JAMA Pediatrics. 2015;169(4):305-6.
39. Sachs JD. Achieving universal health coverage in low-income settings. The Lancet. 2012;380(9845):944-7.
40. Lorch SA, Myers S, Carr B. The Regionalization of Pediatric Health Care. Pediatrics. 2010;126(6):1182-90.
41. Ministry of Health, Government of Kenya. Kenya Health Sector Referral Strategy 2014-2018. In: Division of Emergency and Disaster Risk Management, editor. 2014.
42. World Health Organisation. Quality of care : a process for making strategic choices in health systems. 2006.
43. Clancy CM, Cronin K. Evidence-Based Decision Making: Global Evidence, Local Decisions. Health Affairs. 2005;24(1):151-62.
44. WHO. Systems thinking for health systems strengthening 2009.
45. Mbondji PE, Kebede D, Soumbey-Alley EW, Zielinski C, Kouvividila W, Lusamba-Dikassa P-S. Health information systems in Africa: descriptive analysis of data sources, information products and health statistics. Journal of the Royal Society of Medicine. 2014;107(1 suppl):34-45.
46. Kihuba E, Gathara D, Mwinga S, Mulaku M, Kosgei R, Mogoa W, et al. Assessing the ability of health information systems in hospitals to support evidence-informed decisions in Kenya. Global Health Action. 2014;7.
47. Kerr KA, Norris T, Stockdale R. The strategic management of data quality in healthcare. Health Informatics Journal. 2008;14(4):259-66.
48. Sally K. Stansfield, Julia Walsh, Ndola Prata, Evan T. Information to Improve Decision Making for Health. In: Dean T. Jamison, Joel G. Breman, Anthony R. Measham, George Alleyne, Mariam Claeson, David B. Evans, et al., editors. Disease Control Priorities in Developing Countries: The International Bank for Reconstruction and Development/The World Bank Group.; 2006.
49. Nutley T, Reynolds HW. Improving the use of health data for health system strengthening. 2013. 2013;6.
50. Zwijnenberg NC, Hendriks M, Delnoij DMJ, de Veer AJE, Spreeuwenberg P, Wagner C. Understanding and using quality information for quality improvement: The effect of information presentation. International Journal for Quality in Health Care. 2016.
51. Opiyo N, English M. What clinical signs best identify severe illness in young infants aged 0-59 days in developing countries? A systematic review. Archives of Disease in Childhood. 2011;96(11):1052-9.
52. Wilcox AJ. On the importance and the unimportance of birthweight. International Journal of Epidemiology. 2001;30(6):1233-41.
53. Ministry of Medical Services., Sanitation. MoPH. Comprehensive National Health Policy Framework 2011-2030 (draft): Ministry of Medical Services, Ministry of Public Health & Sanitation; 2011
54. Donabedian Avedis. The seven pillars of quality. Archives of Pathology and Laboratory Medicine. 1990;1990 Nov;114(11):1115-8.
55. Santhakumaran S. How to adjust for case-mix when comparing outcomes across healthcare providers. Archives of disease in childhood - Education & practice edition. 2014;99(2):70-1.
56. Marchant T, Willey B, Katz J, Clarke S, Kariuki S, Kuile F. Neonatal mortality risk associated with preterm birth in East Africa, adjusted by weight for gestational age: individual participant level meta-analysis. PLoS Med. 2012;9.
57. Aluvaala J, Okello D, Murithi G, Wafula L, Wanjala L, Isika N, et al. Delivery outcomes and patterns of morbidity and mortality for neonatal admissions in five Kenyan hospitals. Journal of Tropical Pediatrics. 2015;61(4):255-9.

58. Katz J, Lee ACC, Kozuki N, Lawn JE, Cousens S, Blencowe H, et al. Mortality risk in preterm and small-for-gestational-age infants in low-income and middle-income countries: a pooled country analysis. The Lancet. 2013;382(9890):417-25.
59. Barros FC, Papageorghiou AT, Victora CG, et al. The distribution of clinical phenotypes of preterm birth syndrome: Implications for prevention. JAMA Pediatrics. 2015.
60. The World Bank. Improving the delivery of health services: A guide to choosing strategies. 2011.
61. Watson SI, Arulampalam W, Petrou S, Marlow N, Morgan AS, Draper E, et al. The effects of a one-to-one nurse-to-patient ratio on the mortality rate in neonatal intensive care: a retrospective, longitudinal, population-based study. Archives of Disease in Childhood - Fetal and Neonatal Edition. 2016;101(3):F195-F200.
62. StC Hamilton KE, Redshaw ME, Tarnow-Mordi W. Nurse staffing in relation to risk-adjusted mortality in neonatal care. Archives of Disease in Childhood - Fetal and Neonatal Edition. 2007;92(2):F99-F103.
63. Sherenian M, Profit J, Schmidt B, Suh S, Xiao R, Zupancic JAF, et al. Nurse-to-Patient Ratios and Neonatal Outcomes: A Brief Systematic Review. Neonatology. 2013;104(3):179-83.
64. Ministry of Health, Government of Kenya. Health Sector Human Resources Strategy 2014 – 2018. 2014.
65. West AF, West RR. Clinical decision-making: coping with uncertainty. Postgraduate Medical Journal. 2002;78(920):319-21.
66. Croft P, Altman DG, Deeks JJ, Dunn KM, Hay AD, Hemingway H, et al. The science of clinical practice: disease diagnosis or patient prognosis? Evidence about "what is likely to happen" should shape clinical practice. BMC Medicine. 2015;13(1):1-8.
67. The PROGnosis RESearch Strategy (PROGRESS) Partnership [Available from: http://progress-partnership.org/.
68. Hemingway H, Croft P, Perel P, Hayden JA, Abrams K, Timmis A, et al. Prognosis research strategy (PROGRESS) 1: A framework for researching clinical outcomes. BMJ. 2013;346.
69. Riley RD, Hayden JA, Steyerberg EW, Moons KGM, Abrams K, Kyzas PA, et al. Prognosis Research Strategy (PROGRESS) 2: Prognostic Factor Research. PLoS Med. 2013;10(2):e1001380.
70. Steyerberg EW, Moons KGM, van der Windt DA, Hayden JA, Perel P, Schroter S, et al. Prognosis Research Strategy (PROGRESS) 3: Prognostic Model Research. PLoS Med. 2013;10(2):e1001381.
71. Hingorani AD, Windt DAvd, Riley RD, Abrams K, Moons KGM, Steyerberg EW, et al. Prognosis research strategy (PROGRESS) 4: Stratified medicine research. BMJ : British Medical Journal. 2013;346.
72. Liu L. Global, regional, and national causes of child mortality: an updated systematic analysis for 2010 with time trends since 2000. Lancet. 2012;379(9832):2151 - 61.
73. Were FN, Bwibo NO. Neonatal nutrition and later outcomes of very low birth weight infants at Kenyatta national hospital. African Health Sciences. 2007;7(2):108-14.
74. Mwaniki MK, Nokes DJ, Ignas J, Munywoki P, Ngama M, Newton CR, et al. Emergency triage assessment for hypoxaemia in neonates and young children in a Kenyan hospital: an observational study. Bull World Health Organ. 2009;87(4):263-70.
75. Were FN, Bwibo NO. The contribution of very low birth weight deaths to infant mortality. East African Medical Journal. 2009;86(8):374-7.
76. Mwaniki M, Mathenge A, Gwer S, Mturi N, Bauni E, Newton C, et al. Neonatal seizures in a rural Kenyan District Hospital: aetiology, Incidence and outcome of hospitalization. BMC Medicine. 2010;8(1):16.

77. Mwaniki M, Talbert A, Mturi F, Berkley J, Kager P, Marsh K, et al. Congenital and neonatal malaria in a rural Kenyan district hospital: An eight-year analysis. Malaria Journal. 2010;9(1):313.
78. Marete IK, Wasunna AO, Otieno PA. Clinical risk index for babies (CRIB)II score as a predictor of neonatal mortality among low birth weight babies at Kenyatta National Hospital. East African Medical Journal. 2011;88(1).
79. Kohli-Kochhar R, Omuse G, Revathi G. A ten-year review of neonatal bloodstream infections in a tertiary private hospital in Kenya2011.
80. Yego F, StewartWilliams J, Byles J, Nyongesa P, Aruasa W, Deste C. A retrospective analysis of maternal and neonatal mortality at a teaching and referral hospital in Kenya. Reproductive Health. 2013;10(1):13.
81. Ibinda F, Bauni E, Kariuki SM, Fegan G, Lewa J, Mwikamba M, et al. Incidence and Risk Factors for Neonatal Tetanus in Admissions to Kilifi County Hospital, Kenya. PLOS ONE. 2015;10(4):e0122606.
82. Murphy GAV, Gathara D, Aluvaala J, Mwachiro J, Abuya N, Ouma P, et al. Nairobi Newborn Study: a protocol for an observational study to estimate the gaps in provision and quality of inpatient newborn care in Nairobi City County, Kenya. BMJ Open. 2016;6(12).
83. Barfield WD, Papile L-A, Baley JE, Benitz W, Cummings J, Carlo WA, et al. Levels of Neonatal Care. Pediatrics. 2012;130(3):587-97.
84. Lasswell SM, Barfield WD, Rochat RW, Blackmon L. Perinatal regionalization for very low-birth-weight and very preterm infants. A meta-analysis. JAMA. 2010;304.
85. Neu J. Necrotizing Enterocolitis: The Mystery Goes On. Neonatology. 2014;106(4):289-95.
86. Azzopardi DV, Strohm B, Edwards AD, Dyet L, Halliday HL, Juszczak E, et al. Moderate Hypothermia to Treat Perinatal Asphyxial Encephalopathy. New England Journal of Medicine. 2009;361(14):1349-58.
87. Gluckman PD, Wyatt JS, Azzopardi D, Ballard R, Edwards AD, Ferriero DM, et al. Selective head cooling with mild systemic hypothermia after neonatal encephalopathy: multicentre randomised trial. The Lancet. 2005;365(9460):663-70.
88. Gunn AJ, Gunn TR. The `pharmacology' of neuronal rescue with cerebral hypothermia. Early Human Development. 1998;53(1):19-35.
89. Byass P. Child mortality is (estimated to be) falling. The Lancet. 2016;388(10063):2965-7.
90. Liu L, Oza S, Hogan D, Perin J, Rudan I, Lawn JE, et al. Global, regional, and national causes of child mortality in 2000–13, with projections to inform post-2015 priorities: an updated systematic analysis. The Lancet. 2015;385(9966):430-40.
91. Ho T, Dukhovny D, Zupancic JAF, Goldmann DA, Horbar JD, Pursley DM. Choosing Wisely in Newborn Medicine: Five Opportunities to Increase Value. Pediatrics. 2015.
92. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): The TRIPOD StatementThe TRIPOD Statement. Annals of Internal Medicine. 2015;162(1):55-63.
93. Kleinrouweler CE, Cheong-See FM, Collins GS, Kwee A, Thangaratinam S, Khan KS, et al. Prognostic models in obstetrics: available, but far from applicable. American Journal of Obstetrics & Gynecology. 2015.
94. Vincent J-L, Moreno R. Clinical review: Scoring systems in the critically ill. Critical Care. 2010;14(2):207.
95. Adams ST, Leveson SH. Clinical prediction rules. BMJ. 2012;344:d8312.
96. Bouwmeester W, Zuithoff NPA, Mallett S, Geerlings MI, Vergouwe Y, Steyerberg EW, et al. Reporting and Methods in Clinical Prediction Research: A Systematic Review. PLoS Med. 2012;9(5):e1001221.

97. Collins GS, Omar O, Shanyinde M, Yu L-M. A systematic review finds prediction models for chronic kidney disease were poorly reported and often developed using inappropriate methods. Journal of Clinical Epidemiology. 2013;66(3):268-77.
98. Mallett S, Royston P, Dutton S, Waters R, Altman DG. Reporting methods in studies developing prognostic models in cancer: a review. BMC Med. 2010;8.
99. Collins GS, Mallett S, Omar O, Yu L-M. Developing risk prediction models for type 2 diabetes: a systematic review of methodology and reporting. BMC Medicine. 2011;9(1):103.
100. Steyerberg EW. Clinical Prediction Models. 1 ed: Springer-Verlag New York; 2009. XXVIII, 500 p.
101. Moons KGM, de Groot JAH, Bouwmeester W, Vergouwe Y, Mallett S, Altman DG, et al. Critical Appraisal and Data Extraction for Systematic Reviews of Prediction Modelling Studies: The CHARMS Checklist. PLoS Med. 2014;11(10):e1001744.
102. Dorling JS, Field DJ, Manktelow B. Neonatal disease severity scoring systems. Archives of Disease in Childhood - Fetal and Neonatal Edition. 2005;90(1):F11-F6.
103. Rosenberg RE, Ahmed S, Saha SK, Ahmed ASMNU, Chowdhury MAKA, Law PA, et al. Simplified Age-Weight Mortality Risk Classification for Very Low Birth Weight Infants in Low-Resource Settings. The Journal of Pediatrics. 2008;153(4):519-24.e3.
104. Gray JE, Richardson DK, McCormick MC, Workman-Daniels K, Goldmann DA. Neonatal therapeutic intervention scoring system: a therapy-based severity-of-illness index. Pediatrics. 1992;90(4):561-7.
105. Damen JAAG, Hooft L, Schuit E, Debray TPA, Collins GS, Tzoulaki I, et al. Prediction models for cardiovascular disease risk in the general population: systematic review. BMJ. 2016;353.
106. Collins GS, de Groot JA, Dutton S, Omar O, Shanyinde M, Tajar A, et al. External validation of multivariable prediction models: a systematic review of methodological conduct and reporting. BMC Medical Research Methodology. 2014;14(1):40.
107. Collins GS, Le Manach Y. Statistical Inefficiencies in the Development of a Prediction Model. Anesthesia & Analgesia. 2017;124(3):1011-2.
108. Guyatt GH, Oxman AD, Vist GE, Kunz R, Falck-Ytter Y, Alonso-Coello P, et al. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. BMJ. 2008;336(7650):924-6.
109. Cochrane Bias Methods Group. [Available from: http://methods.cochrane.org/bias/assessing-risk-bias-included-studies.
110. Kleijnen Systematic Reviews Ltd. PROBAST [Available from: http://s371539711.initial-website.co.uk/probast/.
111. Guyatt G, Oxman AD, Akl EA, Kunz R, Vist G, Brozek J, et al. GRADE guidelines: 1. Introduction—GRADE evidence profiles and summary of findings tables. Journal of Clinical Epidemiology. 2011;64(4):383-94.
112. Aluvaala J, Collins GS, Maina M, Berkley JA, English M. A systematic review of neonatal treatment intensity scores and their potential application in low-resource setting hospitals for predicting mortality, morbidity and estimating resource use. Systematic reviews. 2017;6(1):248.
113. Guyatt GH, Oxman AD, Kunz R, Vist GE, Falck-Ytter Y, Schünemann HJ. What is "quality of evidence" and why is it important to clinicians? BMJ. 2008;336(7651):995-8.
114. Guyatt GH, Oxman AD, Kunz R, Atkins D, Brozek J, Vist G, et al. GRADE guidelines: 2. Framing the question and deciding on important outcomes. Journal of Clinical Epidemiology. 2011;64(4):395-400.
115. Balshem H, Helfand M, Schünemann HJ, Oxman AD, Kunz R, Brozek J, et al. GRADE guidelines: 3. Rating the quality of evidence. Journal of Clinical Epidemiology. 2011;64(4):401-6.

116. Neumann I, Santesso N, Akl EA, Rind DM, Vandvik PO, Alonso-Coello P, et al. A guide for health professionals to interpret and use recommendations in guidelines developed with the GRADE approach. Journal of Clinical Epidemiology. 2016;72:45-55.
117. Guyatt GH, Oxman AD, Sultan S, Glasziou P, Akl EA, Alonso-Coello P, et al. GRADE guidelines: 9. Rating up the quality of evidence. Journal of Clinical Epidemiology. 2011;64(12):1311-6.
118. Guyatt GH, Oxman AD, Vist G, Kunz R, Brozek J, Alonso-Coello P, et al. GRADE guidelines: 4. Rating the quality of evidence—study limitations (risk of bias). Journal of Clinical Epidemiology. 2011;64(4):407-15.
119. Guyatt GH, Oxman AD, Kunz R, Woodcock J, Brozek J, Helfand M, et al. GRADE guidelines: 7. Rating the quality of evidence—inconsistency. Journal of Clinical Epidemiology. 2011;64(12):1294-302.
120. Guyatt GH, Oxman AD, Kunz R, Woodcock J, Brozek J, Helfand M, et al. GRADE guidelines: 8. Rating the quality of evidence—indirectness. Journal of Clinical Epidemiology. 2011;64(12):1303-10.
121. Guyatt GH, Oxman AD, Kunz R, Brozek J, Alonso-Coello P, Rind D, et al. GRADE guidelines 6. Rating the quality of evidence—imprecision. Journal of Clinical Epidemiology. 2011;64(12):1283-93.
122. Guyatt GH, Oxman AD, Montori V, Vist G, Kunz R, Brozek J, et al. GRADE guidelines: 5. Rating the quality of evidence—publication bias. Journal of Clinical Epidemiology. 2011;64(12):1277-82.
123. Brożek JL, Akl EA, Jaeschke R, Lang DM, Bossuyt P, Glasziou P, et al. Grading quality of evidence and strength of recommendations in clinical practice guidelines: Part 2 of 3. The GRADE approach to grading quality of evidence about diagnostic tests and strategies. Allergy. 2009;64(8):1109-16.
124. Gopalakrishna G, Mustafa RA, Davenport C, Scholten RJPM, Hyde C, Brozek J, et al. Applying Grading of Recommendations Assessment, Development and Evaluation (GRADE) to diagnostic tests was challenging but doable. Journal of Clinical Epidemiology. 2014;67(7):760-8.
125. Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research electronic data capture (REDCap) ”A metadata-driven methodology and workflow process for providing translational research informatics support. Journal of Biomedical Informatics. 2009;42(2):377-81.
126. Gathara D, Nyamai R, Were F, Mogoa W, Karumbi J, Kihuba E, et al. Moving towards Routine Evaluation of Quality of Inpatient Pediatric Care in Kenya. PLoS ONE. 2015;10(3):e0117048.
127. Mwinga S, Kulohoma C, Mwaniki P, Idowu R, Masasabi J, English M, et al. Quality of surgical care in hospitals providing internship training in Kenya: a cross sectional survey. Tropical Medicine & International Health. 2015;20(2):240-9.
128. R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing.
129. Clark TG, Bradburn MJ, Love SB, Altman DG. Survival Analysis Part I: Basic concepts and first analyses. Br J Cancer. 2003;89(2):232-8.
130. Bradburn MJ, Clark TG, Love SB, Altman DG. Survival Analysis Part II: Multivariate data analysis - an introduction to concepts and methods. Br J Cancer. 2003;89(3):431-6.
131. Therneau T. A Package for Survival Analysis in S. 2.38 ed2015.
132. Pintilie M. An Introduction to Competing Risks Analysis. Revista Española de Cardiología (English Edition). 2011;64(07):599-605.
133. Varadhan R, Weiss CO, Segal JB, Wu AW, Scharfstein D, Boyd C. Evaluating Health Outcomes in the Presence of Competing Risks: A Review of Statistical Methods and Clinical Applications. Medical Care. 2010;48(6):S96-S105.

134. Gray B. cmprsk: Subdistribution Analysis of Competing Risks. 2014.
135. Steyerberg EW, Vergouwe Y. Towards better clinical prediction models: seven steps for development and an ABCD for validation. European Heart Journal. 2014;35(29):1925-31.
136. Steyerberg EW. Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating. New York: Springer; 2009.
137. Harrell Jr FE. Regression Modeling Strategies. R package version 5.1-0 ed2017.
138. Altman DG, Bland JM. Missing data. BMJ. 2007;334(7590):424-.
139. Newgard CD, Lewis RJ. Missing data: How to best account for what is not known. JAMA. 2015;314(9):940-1.
140. Graham JW. Missing data analysis: making it work in the real world. Annual review of psychology. 2009;60:549-76.
141. Little RJAL, Rubin DB. Statistical Analysis with Missing Data, 2nd Edition2002.
142. Abd Hamid IJ, M. Iyen MII, Ibrahim NR, Abd Majid N, Ramli N, Van Rostenberghe H. Randomised controlled trial of single phototherapy with reflecting curtains versus double phototherapy in term newborns with hyperbilirubinaemia. Journal of Paediatrics and Child Health. 2013:n/a-n/a.
143. Aa MNvander, van Leenders GJ, Steyerberg EW, van Rhijn BW, Jobsis AC, Zwarthoff EC, et al. A new system for substaging pT1 papillary bladder cancer: a prognostic evaluation. Hum Pathol. 2005;36.
144. Moons KGM, Altman DG, Reitsma JB, Ioannidis JPA, Macaskill P, Steyerberg EW, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): Explanation and ElaborationThe TRIPOD Statement: Explanation and Elaboration. Annals of Internal Medicine. 2015;162(1):W1-W73.
145. De Cock B, Nieboer D, Van Calster B, Steyerberg E, Vergouwe Y. CalibrationCurves: Calibration performance. 0.1.2 ed2016.
146. Pavlou M, Ambler G, Seaman SR, Guttmann O, Elliott P, King M, et al. How to develop a more accurate risk prediction model when there are few events. BMJ : British Medical Journal. 2015;351.
147. Justice AC, Covinsky KE, Berlin JA. ASsessing the generalizability of prognostic information. Annals of Internal Medicine. 1999;130(6):515-24.
148. Collins GS, Moons KGM. Comparing risk prediction models. BMJ. 2012;344.
149. Vickers AJ, Van Calster B, Steyerberg EW. Net benefit approaches to the evaluation of prediction models, molecular markers, and diagnostic tests. BMJ. 2016;352.
150. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. Med Decis Making. 2006;26.
151. Brown M. DecisionCurve: Calculate and Plot Decision Curves. R package version 1.3 ed2016.
152. Hendriksen JMT, Geersing GJ, Moons KGM, de Groot JAH. Diagnostic and prognostic prediction models. Journal of Thrombosis and Haemostasis. 2013;11:129-41.
153. Dorling JS, Field DJ. Value and validity of neonatal disease severity scoring systems. Archives of Disease in Childhood - Fetal and Neonatal Edition. 2008;93(2):F80-F2.
154. Virginia Apgar. A proposal for a new method of evaluation of the newborn infant. Current researches in anesthesia and analgesia. 1953.
155. Silverman WA, Andersen DH. A controlled clinical trial of effects of water mist on obstructive respiratory signs, death rate and necropsy findings among premature infants. Pediatrics. 1956;17(1):1-10.
156. Bell MJ, Ternberg JL, Feigin RD, Keating JP, Marshall R, Barton L, et al. Neonatal Necrotizing Enterocolitis: Therapeutic Decisions Based upon Clinical Staging. Annals of Surgery. 1978;187(1):1-7.
157. Kliegman RM, Walsh MC. Neonatal necrotizing enterocolitis: Pathogenesis, classification, and spectrum of illness. Current Problems in Pediatrics. 1987;17(4):219-88.

158. Papile L-A, Burstein J, Burstein R, Koffler H. Incidence and evolution of subependymal and intraventricular hemorrhage: A study of infants with birth weights less than 1,500 gm. The Journal of Pediatrics. 1978;92(4):529-34.
159. Sarnat HB, Sarnat MS. Neonatal encephalopathy following fetal distress: A clinical and electroencephalographic study. Archives of Neurology. 1976;33(10):696-705.
160. Iliodromiti S, Mackay DF, Smith GCS, Pell JP, Nelson SM. Apgar score and the risk of cause-specific infant mortality: a population-based cohort study. The Lancet. 2014;384(9956):1749-55.
161. Dubowitz LMS, Dubowitz V, Goldberg C. Clinical assessment of gestational age in the newborn infant. The Journal of Pediatrics. 1970;77(1):1-10.
162. Ballard JL, Novak KK, Driver M. A simplified score for assessment of fetal maturation of newly born infants. The Journal of Pediatrics. 1979;95(5, Part 1):769-74.
163. Ballard JL, Khoury JC, Wedig K, Wang L, Eilers-Walsman BL, Lipp R. New Ballard Score, expanded to include extremely premature infants. The Journal of Pediatrics. 1991;119(3):417-23.
164. Rodwell RL, Leslie AL, Tudehope DI. Early diagnosis of neonatal sepsis using a hematologic scoring system. The Journal of Pediatrics. 1988;112(5):761-7.
165. Geersing GJ, Bouwmeester W, Zuithoff P, Spijker R, Leeflang M, Moons K. Search filters for finding prognostic and diagnostic prediction studies in medline to enhance systematic reviews. PLoS One. 2012;7.
166. Fleisher BE, Murthy L, Lee S, Constantinou JC, Benitz WE, Stevenson DK. Neonatal Severity of Illness Scorinng Systems: A Comparison. Clinical Pediatrics. 1997;36(4):223-7.
167. Marcin JP, Pollack MM. Review of the methodologies and applications of scoring systems in neonatal and pediatric intensive care. Pediatric Critical Care Medicine. 2000;1(1):20-7.
168. Patrick SW, Schumacher RE, Davis MM. Methods of Mortality Risk Adjustment in the NICU: A 20-Year Review. Pediatrics. 2013;131(Supplement 1):S68-S74.
169. Sacco Casamassima MG, Salazar JH, Papandria D, Fackler J, Chrouser K, Boss EF, et al. Use of risk stratification indices to predict mortality in critically ill children. European Journal of Pediatrics. 2014;173(1):1-13.
170. Agrawal PK, Agrawal S, Ahmed S, Darmstadt GL, Williams EK, Rosen HE, et al. Effect of knowledge of community health workers on essential newborn health care: a study from rural India. Health Policy and Planning.27(2):115-26.
171. Tarnow-Mordi WO. What is the role of neonatal organ dysfunction and illness severity scores in therapeutic studies in sepsis? Pediatric Critical Care Medicine. 2005;6(3):S135-S7.
172. Pollack MM, Koch MA, Bartel DA, Rapoport I, Dhanireddy R, El-Mohandes AAE, et al. A Comparison of Neonatal Mortality Risk Prediction Models in Very Low Birth Weight Infants. Pediatrics. 2000;105(5):1051-7.
173. Medlock S, Ravelli ACJ, Tamminga P, Mol BWM, Abu-Hanna A. Prediction of Mortality in Very Premature Infants: A Systematic Review of Prediction Models. PLoS ONE. 2011;6(9):e23441.
174. Richardson DK, Gray JE, McCormick MC, Workman K, Goldmann DA. Score for Neonatal Acute Physiology: A Physiologic Severity Index for Neonatal Intensive Care. Pediatrics. 1993;91(3):617-23.
175. Richardson DK, Corcoran JD, Escobar GJ, Lee SK. SNAP-II and SNAPPE-II: Simplified newborn illness severity and mortality risk scores. The Journal of Pediatrics. 2001;138(1):92-100.
176. Richardson DK, Phibbs CS, Gray JE, McCormick MC, Workman-Daniels K, Goldmann DA. Birth Weight and Illness Severity: Independent Predictors of Neonatal Mortality. Pediatrics. 1993;91(5):969-75.

177. Sinkin RA, Cox C, Phelps DL. Predicting Risk for Bronchopulmonary Dysplasia: Selection Criteria for Clinical Trials. Pediatrics. 1990;86(5):728-36.
178. Brazy JE, Eckerman CO, Oehler JM, Goldstein RF, O'Rand AM. Nursery Neurobiologic Risk Score: Important factors in predicting outcome in very low birth weight infants. The Journal of Pediatrics. 1991;118(5):783-92.
179. Lee SK, Zupancic JAF, Pendray M, Thiessen P, Schmidt B, Whyte R, et al. Transport risk index of physiologic stability: A practical system for assessing infant transport care. The Journal of Pediatrics. 2001;139(2):220-6.
180. Lee SK, Aziz K, Dunn M, Clarke M, Kovacs L, Ojah C, et al. Transport Risk Index of Physiologic Stability, Version II (TRIPS-II): A Simple and Practical Neonatal Illness Severity Score. Amer J Perinatol. 2013;30(05):395-400.
181. Zupancic JAF, Richardson DK, Horbar JD, Carpenter JH, Lee SK, Escobar GJ. Revalidation of the Score for Neonatal Acute Physiology in the Vermont Oxford Network. Pediatrics. 2007;119(1):e156-e63.
182. The International Neonatal N. The CRIB (clinical risk index for babies) score: a tool for assessing initial neonatal risk and comparing performance of neonatal intensive care units. The Lancet. 1993;342(8865):193-8.
183. Parry G, Tucker J, Tarnow-Mordi W. CRIB II: an update of the clinical risk index for babies score. The Lancet. 2003;361(9371):1789-91.
184. Maier RF, Rey M, Metze BC, Obladen M. Comparison of mortality risk: a score for very low birthweight infants. Archives of Disease in Childhood - Fetal and Neonatal Edition. 1997;76(3):F146-F51.
185. Horbar JD, Onstad L, Wright E, Health TNIOC, Network HDNR. Predicting mortality risk for infants weighing 501 to 1500 grams at birth: A National Institutes of Health Neonatal Research Network report. Critical Care Medicine. 1993;21(1):12-8.
186. Tyson JE, Parikh NA, Langer J, Green C, Higgins RD. Intensive Care for Extreme Prematurity — Moving beyond Gestational Age. New England Journal of Medicine. 2008;358(16):1672-81.
187. Evans N, Hutchinson J, Simpson JM, Donoghue D, Darlow B, Henderson-Smart D. Prenatal predictors of mortality in very preterm infants cared for in the Australian and New Zealand Neonatal Network. Archives of Disease in Childhood - Fetal and Neonatal Edition. 2007;92(1):F34-F40.
188. Marshall G, Tapia JL, D'Apremont I, Grandi C, Barros C, Alegria A, et al. A New Score for Predicting Neonatal Very Low Birth Weight Mortality Risk in the NEOCOSUR South American Network. J Perinatol. 2005;25(9):577-82.
189. Janota J, Stranák Z, Statecá B, Dohnalová A, Šípek A, Šimák J. Characterization of multiple organ dysfunction syndrome in very low birthweight infants: a new sequential scoring system. Shock. 2001;15(5):348-52.
190. Moher D, Liberati A, Tetzlaff J, Altman DG. Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. Journal of Clinical Epidemiology. 2009;62(10):1006-12.
191. Shah PS, Mirea L, Ng E, Solimano A, Lee SK. Association of unit size, resource utilization and occupancy with outcomes of preterm infants. J Perinatol. 2015;35(7):522-9.
192. Davies MRQ. The need for a universal method of quantifying severity of illness to allow accurate analysis of the results of treatment in neonatal surgical cases. Pediatr Surg Int. 1995;10(5-6):305-8.
193. Eriksson M, Bodin L, Finnström O, Schollin J. Can severity-of-illness indices for neonatal intensive care predict outcome at 4 years of age? Acta Pædiatrica. 2002;91(10):1093-100.

194. Zupancic JA, Richardson DK. Characterization of neonatal personnel time inputs and prediction from clinical variables--a time and motion study. J Perinatol. 2002;22(8):658-63.
195. Mendes I, Carvalho M, Almeida RT, Moreira ME. Use of technology as an evaluation tool of clinical care in preterm newborns. Jornal de pediatria. 2006;82(5):371-6.
196. Rojas JG, Henao-Murillo NA, Quirós-Jaramillo A. A tool for calculating the nursing staff at neonatal intensive care units. Aquichán. 2011;11:126-39.
197. Oygur N, Ongun H, Saka O. Risk prediction using a neonatal therapeutic intervention scoring system in VLBW and ELBW preterm infants. Pediatrics International. 2012;54(4):496-500.
198. Wu P-L, Lee W-T, Lee P-L, Chen H-L. Predictive Power of Serial Neonatal Therapeutic Intervention Scoring System Scores for Short-term Mortality in Very-low-birth-weight Infants. Pediatrics & Neonatology. 2015;56(2):108-13.
199. Georgieff MKM, Mills MMR, Bhatt PB. Validation of two scoring systems which assess the degree of physiologic instability in critically ill newborn infants. Critical Care Medicine. 1989;17(1):17-21.
200. Cullen DJM, Civetta JMM, Briggs BAM, Ferrara LCR. Therapeutic intervention scoring system: a method for quantitative comparison of patient care. Critical Care Medicine March/April. 1974;2(2):57-60.
201. Keene ARR, Cullen DJM. Therapeutic Intervention Scoring System: Update 1983. Critical Care Medicine. 1983;11(1):1-3.
202. Vergouwe Y, Steyerberg EW, Eijkemans MJC, Habbema JDF. Substantial effective sample sizes were required for external validation studies of predictive logistic regression models. Journal of Clinical Epidemiology. 2005;58(5):475-83.
203. Collins GS, Ogundimu EO, Altman DG. Sample size considerations for the external validation of a multivariable prognostic model: a resampling study. Statistics in Medicine. 2016;35(2):214-26.
204. Van Calster B, Nieboer D, Vergouwe Y, De Cock B, Pencina MJ, Steyerberg EW. A calibration hierarchy for risk models was defined: from utopia to empirical data. Journal of Clinical Epidemiology. 2016;74:167-76.
205. Van Calster B, Nieboer D, Vergouwe Y, De Cock B, Pencina MJ, Steyerberg EW. A calibration hierarchy for risk models was defined: from utopia to empirical data. Journal of Clinical Epidemiology.74:167-76.
206. Altman DG, Royston P. What do we mean by validating a prognostic model? Statistics in Medicine. 2000;19(4):453-73.
207. Altman DG, Vergouwe Y, Royston P, Moons KGM. Prognosis and prognostic research: validating a prognostic model. BMJ. 2009;338.
208. Shmueli G. To Explain or to Predict? Statist Sci. 2010:289-310.
209. Marks FE. Refining a Classification System for Fiscal and Staffing Management. Journal of Nursing Administration. 1987;17(1):39.
210. Cullen DJMM, Nemeskal ARR, Zaslavsky AMP. Intermediate TISS: A new Therapeutic Intervention Scoring System for non-ICU patients. Critical Care Medicine. 1994;22(9):1406-11.
211. Gemke RJ, Bonsel GJ, McDonnell J, van Vught AJ. Patient characteristics and resource utilisation in paediatric intensive care. Archives of Disease in Childhood. 1994;71(4):291-6.
212. Reis Miranda DMD, de Rijk AB, Schaufeli WP. Simplified Therapeutic Intervention Scoring System: The TISS-28 items--Results from a multicenter study. Critical Care Medicine. 1996;24(1):64-73.
213. Sainani KL. Explanatory Versus Predictive Modeling. PM&R. 2014;6(9):841-4.

214. Kozma CM, Reeder CE, Schulz RM. Economic, clinical, and humanistic outcomes: a planning model for pharmacoeconomic research. Clinical therapeutics. 1993;15(6):1121-32; discussion 0.
215. Velentgas P, Dreyer NA, Nourjah P, Smith SR, Torchia MM. Developing a Protocol for Observational Comparative Effectiveness Research: A User's Guide. Agency for Healthcare Research and Quality; 2013.
216. Kieny MP, Bekedam H, Dovlo D, Fitzgerald J, Habicht J, Harrison G, et al. Strengthening health systems for universal health coverage and sustainable development. Bull World Health Organ. 2017;95(7):537-9.
217. Gargon E, Gurung B, Medley N, Altman DG, Blazeby JM, Clarke M, et al. Choosing important health outcomes for comparative effectiveness research: a systematic review. PLoS One. 2014;9(6):e99111.
218. Every Newborn: an action plan to end preventable deaths. Geneva: World Health Organization; 2014.
219. World Health Organization. Global Reference List of 100 Core Health Indicators. 2015.
220. Hayden JA, Côté P, Steenstra IA, Bombardier C. Identifying phases of investigation helps planning, appraising, and applying the results of explanatory prognosis studies. Journal of Clinical Epidemiology. 2008;61(6):552-60.
221. Stewart K, Choudry MI, Buckingham R. Learning from hospital mortality. Clinical Medicine. 2016;16(6):530-4.
222. Rachel MM, Stewart MW. Establishing a Mortality Review Process. Journal of Nursing Care Quality. 2009;24(3).
223. Saleem S, McClure Elizabeth M, Goudar Shivaprasad S, Patel A, Fabian Esamai, Ana Garces, et al. A prospective study of maternal, fetal and neonatal deaths in low- and middle-income countries. Bulletin of the World Health Organization. 2014;92:605–12.
224. Baqui AH, Mitra DK, Begum N, Hurt L, Soremekun S, Edmond K, et al. Neonatal mortality within 24 hours of birth in six low- and lower-middle-income countries. Bull World Health Organ. 2016;94(10):752-8b.
225. Althabe F, Belizan J, Mazzoni A, Berrueta M, Hemingway-Foday J, Koso-Thomas M, et al. Antenatal corticosteroids trial in preterm births to increase Neonatal survival in developing countries: Study protocol. Reproductive Health.9(1):22.
226. Kenya Demographic and Health Survey 2014. Nairobi: KNBS, MOH, NACC, KEMRI and NCPD; 2015.
227. Health Information Systems Programme. District Health Information System. Version 2 ed.
228. Thomson J, Schaefer M, Caminoa B, Kahindi D, Hurtado N. Improved Neonatal Mortality at a District Hospital in Aweil, South Sudan. J Trop Pediatr. 2017;63(3):189-95.
229. Ndombo PK, Ekei QM, Tochie JN, Temgoua MN, Angong FTE, Ntock FN, et al. A cohort analysis of neonatal hospital mortality rate and predictors of neonatal mortality in a sub-urban hospital of Cameroon. Italian Journal of Pediatrics. 2017;43(1):52.
230. Abdallah Y, Namiiro F, Mugalu J, Nankunda J, Vaucher Y, McMillan D. Is facility based neonatal care in low resource setting keeping pace? A glance at Uganda's National Referral Hospital. Afr Health Sci. 2016;16(2):347-55.
231. Kokeb M, Desta T. Institution Based Prospective Cross-Sectional Study on Patterns of Neonatal Morbidity at Gondar University Hospital Neonatal Unit, North-West Ethiopia. Ethiopian journal of health sciences. 2016;26(1):73-9.
232. Ndelema B, Van den Bergh R, Manzi M, van den Boogaard W, Kosgei RJ, Zuniga I, et al. Low-tech, high impact: care for premature neonates in a district hospital in Burundi. A way forward to decrease neonatal mortality. BMC Research Notes. 2016;9(1):1-8.

233. Cavicchiolo ME, Lanzoni P, Wingi MO, Pizzol D, Daverio M, Da Dalt L, et al. Reduced neonatal mortality in a regional hospital in Mozambique linked to a Quality Improvement intervention. BMC Pregnancy and Childbirth. 2016;16(1):366.
234. Tochie JN, Choukem S-P, Langmia RN, Barla E, Koki-Ndombo P. Neonatal respiratory distress in a reference neonatal unit in Cameroon: an analysis of prevalence, predictors, etiologies and outcomes. Pan Afr Med J. 2016;24.
235. Ballot DE, Chirwa T, Ramdin T, Chirwa L, Mare I, Davies VA, et al. Comparison of morbidity and mortality of very low birth weight infants in a Central Hospital in Johannesburg between 2006/2007 and 2013. BMC Pediatr. 2015;15:20.
236. Okomo UA, Dibbasey T, Kassama K, Lawn JE, Zaman SMA, Kampmann B, et al. Neonatal admissions, quality of care and outcome: 4 years of inpatient audit data from The Gambia's teaching hospital. Paediatrics and International Child Health. 2015;35(3):252-64.
237. Hedstrom A, Ryman T, Otai C, Nyonyintono J, McAdams RM, Lester D, et al. Demographics, clinical characteristics and neonatal outcomes in a rural Ugandan NICU. BMC Pregnancy Childbirth. 2014;14:327.
238. Ahlsén AK, Spong E, Kafumba N, Kamwendo F, Wolff K. Born too small: who survives in the public hospitals in Lilongwe, Malawi? Archives of Disease in Childhood - Fetal and Neonatal Edition. 2015;100(2):F150-F4.
239. Chiabi A, Takou V, Mah E, Nguefack S, Siyou H, Takou V. Risk factors for neonatal mortality at the Yaounde Gynaeco-obstetric and pediatric hospital, Cameroon. Iran J Pediatr. 2014;24.
240. Ekwochi U, Ndu I, Nwokoye I, Ezenwosu O, Amadi O, Osuorah D. Pattern of morbidity and mortality of newborns admitted into the sick and special care baby unit of Enugu State University Teaching Hospital, Enugu state. Nigerian Journal of Clinical Practice. 2014;17(3):346-51.
241. Mah M, Chiabi A, Tchokoteu P, Nguefack S, Bogne J, Siyou H. Neonatal mortality in a referral hospital in Cameroon over a seven year period: trends, associated factors and causes. Afr Health Sci. 2014;14.
242. Zash RM, Ajose-Popoola O, Stordal K, Souda S, Ogwu A, Dryden-Peterson S, et al. Risk factors for mortality among human immunodeficiency virus-exposed and unexposed infants admitted to a neonatal intensive care unit in Botswana. Journal of Paediatrics and Child Health.n/a-n/a.
243. Zuniga I, Van den Bergh R, Ndelema B, Bulckaert D, Manzi M, Lambert V, et al. Characteristics and mortality of neonates in an emergency obstetric and neonatal care facility, rural Burundi. Public health action. 2013;3(4):276-81.
244. Mmbaga B, Lie R, Olomi R, Mahande M, Kvale G, Daltveit A. Cause-specific neonatal mortality in a neonatal care unit in Northern Tanzania: a registry based cohort study. BMC Pediatrics. 2012;12(1):116.
245. Ersdal HL, Mduma E, Svensen E, Perlman J. Birth Asphyxia: A Major Cause of Early Neonatal Mortality in a Tanzanian Rural Hospital. Pediatrics.
246. Pepler PT, Uys DW, Nel DG. Predicting mortality and length-of-stay for neonatal admissions to private hospital neonatal intensive care units: a Southern African retrospective study. Afr Health Sci. 2012;12(2):166-73.
247. Shah S, Zemichael O, Meng HD. Factors associated with mortality and length of stay in hospitalised neonates in Eritrea, Africa: a cross-sectional study. BMJ Open. 2012;2(5).
248. Onwuanaku CA, Okolo SN, Ige KO, Okpe SE, Toma BO. The effects of birth weight and gender on neonatal mortality in north central Nigeria. BMC Res Notes. 2011;4:562.
249. Ballot DE, Chirwa TF, Cooper PA. Determinants of survival in very low birth weight neonates in a public sector hospital in Johannesburg. BMC Pediatrics. 2010;10(1):30.

250. Omoigberale A, Sadoh W, Nwaneri D. A 4 year review of neonatal outcome at the University of Benin teaching hospital, Benin City. Nigerian Journal of Clinical Practice. 2010;13(3):321-5.
251. Myhre J, Immaculate M, Okeyo B, Anand M, Omoding A, Myhre L, et al. Effect of Treatment of Premature Infants with Respiratory Distress Using Low-cost Bubble CPAP in a Rural African Hospital. J Trop Pediatr. 2016;62(5):385-9.
252. Hinchliffe SR, Seaton SE, Lambert PC, Draper ES, Field DJ, Manktelow BN. Modelling Time to Death or Discharge in Neonatal Care: An Application of Competing Risks. Paediatric and Perinatal Epidemiology. 2013;27(4):426-33.
253. Duke T, Yano E, Hutchinson A, Hwaihwanje I, Aipit J, Tovilu M, et al. Large-scale data reporting of paediatric morbidity and mortality in developing countries: it can be done. Arch Dis Child. 2016;101(4):392-7.
254. Gathara D, Malla L, Ayieko P, Karuri S, Nyamai R, Irimu G, et al. Variation in and risk factors for paediatric inpatient all-cause mortality in a low income setting: data from an emerging clinical information network. BMC Pediatr. 2017;17(1):99.
255. Alboukadel Kassambara, Marcin Kosinski. survminer: Drawing Survival Curves using 'ggplot2'. R package version 0.4.0 ed2017.
256. Haumont D, NguyenBa C, Modi N. eNewborn: The Information Technology Revolution and Challenges for Neonatal Networks. Neonatology. 2017;111(4):388-97.
257. The Australia and New Zealand Neonatal Network [Available from: https://www.anznn.net.
258. Chow SSW, Le Marsney R, Creighton P, Kander V, Haslam R, Lui K. Report of the Australian and New Zealand Neonatal Network 2015. 2017.
259. Prakesh S, Eugene W. Y, Priscilla C, Committee MotARR. The Canadian Neonatal Network Annual Report 2015. 2015.
260. Horbar JD, Edwards EM, Greenberg LT, Morrow KA, Soll RF, Buus-Frank ME, et al. Variation in Performance of Neonatal Intensive Care Units in the United States. JAMA Pediatr. 2017;171(3):e164396.
261. Lee AH, Fung WK, Fu B. Analyzing hospital length of stay: mean or median regression? Med Care. 2003;41(5):681-6.
262. Seaton SE, Barker L, Jenkins D, Draper ES, Abrams KR, Manktelow BN. What factors predict length of stay in a neonatal unit: a systematic review. BMJ Open. 2016;6(10).
263. Austin PC, Lee DS, Fine JP. Introduction to the Analysis of Survival Data in the Presence of Competing Risks. Circulation. 2016;133(6):601-9.
264. van Walraven C, McAlister FA. Competing risk bias was common in Kaplan–Meier risk estimates published in prominent medical journals. Journal of Clinical Epidemiology. 2016;69(Supplement C):170-3.e8.
265. Altman DG, Royston P. The cost of dichotomising continuous variables. BMJ. 2006;332.
266. Blencowe H, Cousens S, Oestergaard M, Chou D, Moller A-B, Narwal R, et al. 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 implications. Lancet. 2012;379(9832):2162 - 72.
267. Lee AC, Katz J, Blencowe H, Cousens S, Kozuki N, Vogel JP. National and regional estimates of term and preterm babies born small for gestational age in 138 low-income and middle-income countries in 2010. Lancet Glob Health. 2013;1.
268. Longo S, Bollani L, Decembrino L, Di Comite A, Angelini M, Stronati M. Short-term and long-term sequelae in intrauterine growth retardation (IUGR). The Journal of Maternal-Fetal & Neonatal Medicine. 2013;26(3):222-5.
269. Steyerberg EW, Moons KGM, van der Windt DA, Hayden JA, Perel P, Schroter S, et al. Prognosis Research Strategy (PROGRESS) 3: prognostic model research. PLoS Med. 2013;10.

270. World Health Organization. WHO Guidelines Approved by the Guidelines Review Committee. Geneva: World Health Organization Copyright (c) World Health Organization 2013.; 2013.
271. Groenwold RH, Moons KG, Pajouheshnia R, Altman DG, Collins GS, Debray TP, et al. Explicit inclusion of treatment in prognostic modeling was recommended in observational and randomized settings. J Clin Epidemiol. 2016;78:90-100.
272. Steyerberg EW, Harrell FE. Prediction models need appropriate internal, internal–external, and external validation. Journal of Clinical Epidemiology. 2016;69(Supplement C):245-7.
273. Alkema L, Chao F, You D, Pedersen J, Sawyer CC. National, regional, and global sex ratios of infant, child, and under-5 mortality and identification of countries with outlying ratios: a systematic assessment. The Lancet Global Health. 2014;2(9):e521-e30.
274. Ogundimu EO, Altman DG, Collins GS. Adequate sample size for developing prediction models is not simply related to events per variable. Journal of Clinical Epidemiology. 2016;76:175-82.
275. Janssen KJM, Moons KGM, Kalkman CJ, Grobbee DE, Vergouwe Y. Updating methods improved the performance of a clinical prediction model in new patients. Journal of Clinical Epidemiology. 2008;61(1):76-86.
276. Hosmer DW, Lemeshow S, Sturdivant RX. Applied Logistic Regression. New York, UNITED STATES: John Wiley & Sons, Incorporated; 2013.
277. Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, et al. Assessing the performance of prediction models: a framework for traditional and novel measures. Epidemiology. 2010;21.
278. Kattan MW, Harrell FE. We should not be so quick to abandon the use of domain experts and full models (letter commenting: J Clin Epidemiol. 2015;71C:76--85.). Journal of Clinical Epidemiology. 2016;75(Supplement C):131.
279. Steyerberg EW, Bleeker SE, Moll HA, Grobbee DE, Moons KGM. Internal and external validation of predictive models: A simulation study of bias and precision in small samples. Journal of Clinical Epidemiology. 2003;56(5):441-7.
280. Toll DB, Janssen KJ, Vergouwe Y, Moons KG. Validation, updating and impact of clinical prediction rules: a review. J Clin Epidemiol. 2008;61(11):1085-94.
281. English M, Ayieko P, Nyamai R, Were F, Githanga D, Irimu G. What do we think we are doing? How might a clinical information network be promoting implementation of recommended paediatric care practices in Kenyan hospitals? Health Res Policy Syst. 2017;15(1):4.
282. Ayieko P, Ogero M, Makone B, Julius T, Mbevi G, Nyachiro W, et al. Characteristics of admissions and variations in the use of basic investigations, treatments and outcomes in Kenyan hospitals within a new Clinical Information Network. Arch Dis Child. 2016;101(3):223-9.
283. Tuti T, Bitok M, Malla L, Paton C, Muinga N, Gathara D, et al. Improving documentation of clinical care within a clinical information network: an essential initial step in efforts to understand and improve care in Kenyan hospitals. BMJ global health. 2016;1(1):e000028.
284. Collins GS, Altman DG. An independent and external validation of QRISK2 cardiovascular disease risk score: a prospective open cohort study. BMJ. 2010;340.
285. Hippisley-Cox J, Coupland C, Vinogradova Y, Robson J, May M, Brindle P. Derivation and validation of QRISK, a new cardiovascular disease risk score for the United Kingdom: prospective open cohort study. Bmj. 2007;335(7611):136.
286. Mwaniki MK, Talbert AW, Njuguna P, English M, Were E, Lowe BS, et al. Clinical indicators of bacterial meningitis among neonates and young infants in rural Kenya. BMC Infect Dis. 1186;11(301):1471-2334.

287. Ahmed MI, Iqbal M, Hussain N. A structured approach to the assessment of a floppy neonate. Journal of pediatric neurosciences. 2016;11(1):2-6.
288. Gachau S, Ayieko P, Gathara D, Mwaniki P, Ogero M, Akech S, et al. Does audit and feedback improve the adoption of recommended practices? Evidence from a longitudinal observational study of an emerging clinical network in Kenya. BMJ global health. 2017;2(4).
289. White IR, Royston P, Wood AM. Multiple imputation using chained equations: issues and guidance for practice. Stat Med. 2011;30.
290. Harrell Jr FE, with contributions from Charles Dupont and many others. Hmisc: Harrell Miscellaneous.R package version 4.0-3. 2017.
291. Rubin DB. Multiple Imputation for Nonresponse in Surveys. New York: John Wiley and Sons; 2004.
292. Moons KG, Kengne AP, Woodward M, Royston P, Vergouwe Y, Altman DG, et al. Risk prediction models: I. Development, internal validation, and assessing the incremental value of a new (bio) marker. Heart. 2012;98.
293. Karim MN, Reid CM, Tran L, Cochrane A, Billah B. Variable selection methods for multiple regressions influence the parsimony of risk prediction models for cardiac surgery. J Thorac Cardiovasc Surg. 2017;153(5):1128-35.e3.
294. Tzoulaki I, Liberopoulos G, Ioannidis JP. Use of reclassification for assessment of improved prediction: an empirical evaluation. Int J Epidemiol. 2011;40(4):1094-105.
295. Pepe MS, Fan J, Feng Z, Gerds T, Hilden J. The Net Reclassification Index (NRI): a Misleading Measure of Prediction Improvement Even with Independent Test Data Sets. Statistics in biosciences. 2015;7(2):282-95.
296. Pepe MS, Janes H, Li CI. Net risk reclassification p values: valid or misleading? J Natl Cancer Inst. 2014;106(4):dju041.
297. Cook NR. Clinically relevant measures of fit? A note of caution. Am J Epidemiol. 2012;176(6):488-91.
298. Hilden J, Gerds TA. A note on the evaluation of novel biomarkers: do not rely on integrated discrimination improvement and net reclassification index. Stat Med. 2014;33(19):3405-14.
299. Reilly BM, Evans AT. Translating clinical research into clinical practice: impact of using prediction rules to make decisions. Ann Intern Med. 2006;144(3):201-9.
300. Ebell M. AHRQ White Paper: Use of clinical decision rules for point-of-care decision support. Med Decis Making. 2010;30(6):712-21.
301. Richardson S, Khan S, McCullagh L, Kline M, Mann D, McGinn T. Healthcare provider perceptions of clinical prediction rules. BMJ Open. 2015;5(9).
302. Pluddemann A, Wallace E, Bankhead C, Keogh C, Van der Windt D, Lasserson D, et al. Clinical prediction rules in practice: review of clinical guidelines and survey of GPs. The British journal of general practice : the journal of the Royal College of General Practitioners. 2014;64(621):e233-42.
303. Vickers AJ, Cronin AM, Elkin EB, Gonen M. Extensions to decision curve analysis, a novel method for evaluating diagnostic tests, prediction models and molecular markers. BMC Medical Informatics and Decision Making. 2008;8(1):1-17.
304. Sullivan LM, Massaro JM, D'Agostino RB. Presentation of multivariate data for clinical use: The Framingham Study risk score functions. Statistics in Medicine. 2004;23(10):1631-60.
305. Kerr KF, Brown MD, Zhu K, Janes H. Assessing the Clinical Impact of Risk Prediction Models With Decision Curves: Guidance for Correct Interpretation and Appropriate Use. J Clin Oncol. 2016;34(21):2534-40.

306. Collins GS, Altman DG. Predicting the 10 year risk of cardiovascular disease in the United Kingdom: independent and external validation of an updated version of QRISK2. Bmj. 2012;344:e4181.
307. Rousson V, Zumbrunn T. Decision curve analysis revisited: overall net benefit, relationships to ROC curve analysis, and application to case-control studies. BMC Med Inform Decis Mak. 2011;11:45.
308. Campbell DJ. The clinical utility curve: a proposal to improve the translation of information provided by prediction models to clinicians. BMC Research Notes. 2016;9(1):219.
309. Akanbi MO, Ocheke AN, Agaba PA, Daniyam CA, Agaba EI, Okeke EN, et al. Use of Electronic Health Records in sub-Saharan Africa: Progress and challenges. Journal of medicine in the tropics. 2012;14(1):1-6.
310. UNICEF W, World Bank, UN-DESA Population Division. Levels and trends in child mortality report 2017: Estimates Developed by the UN Inter-agency Group for Child Mortality Estimation. 2017.
311. Bhutta ZA, Das JK, Bahl R, Lawn JE, Salam RA, Paul VK. Lancet Newborn Interventions Review Group; Lancet Every Newborn Study Group. Can available interventions end preventable deaths in mothers, newborn babies, and stillbirths, and at what cost? Lancet. 2014;384.
312. Adamkin DH. Neonatal hypoglycemia. Seminars in Fetal and Neonatal Medicine. 2017;22(1):36-41.
313. Tuti T, Bitok M, Paton C, Makone B, Malla L, Muinga N, et al. Innovating to enhance clinical data management using non-commercial and open source solutions across a multi-center network supporting inpatient pediatric care and research in Kenya. Journal of the American Medical Informatics Association. 2016;23(1):184-92.
314. Thomas J, Ayieko P, Ogero M, Gachau S, Makone B, Nyachiro W, et al. Blood Transfusion Delay and Outcome in County Hospitals in Kenya. The American Journal of Tropical Medicine and Hygiene. 2016.
315. Moxon SG, Ruysen H, Kerber KJ, Amouzou A, Fournier S, Grove J, et al. Count every newborn; a measurement improvement roadmap for coverage data. BMC Pregnancy and Childbirth. 2015;15(2):1-23.
316. Webbe J, Brunton G, Ali S, Duffy JM, Modi N, Gale C. Developing, implementing and disseminating a core outcome set for neonatal medicine. BMJ Paediatrics Open. 2017;1(1).
317. Flenady V, Wojcieszek AM, Ellwood D, Leisher SH, Erwich JJHM, Draper ES, et al. Classification of causes and associated conditions for stillbirths and neonatal deaths. Seminars in Fetal and Neonatal Medicine. 2017;22(3):176-85.
318. Fitchett EJA, Seale AC, Vergnano S, Sharland M, Heath PT, Saha SK, et al. Strengthening the Reporting of Observational Studies in Epidemiology for Newborn Infection (STROBE-NI): an extension of the STROBE statement for neonatal infection research. The Lancet Infectious Diseases.16(10):e202-e13.
319. Williamson PR, Altman DG, Blazeby JM, Clarke M, Gargon E. The COMET (Core Outcome Measures in Effectiveness Trials) Initiative. Trials. 2011;12(Suppl 1):A70-A.
320. Bhutta ZA, Das JK, Bahl R, Lawn JE, Salam RA, Paul VK. Can available interventions end preventable deaths in mothers, newborn babies, and stillbirths, and at what cost? Lancet. 2014;384.
321. American Academy of Pediatrics Committee on Fetus and Newborn. Levels of Neonatal Care. Pediatrics. 2012;130(3):587-97.
322. Murphy GAV, Waters D, Ouma PO, Gathara D, Shepperd S, Snow RW, et al. Estimating the need for inpatient neonatal services: an iterative approach employing evidence and expert consensus to guide local policy in Kenya. BMJ global health. 2017;2(4).

323. Phibbs CS, Baker LC, Caughey AB, Danielsen B, Schmitt SK, Phibbs RH. Level and volume of neonatal intensive care and mortality in very-low-birth-weight infants. N Engl J Med. 2007;356.
324. Persson LÅ. Bridging the quality chasm in maternal, newborn, and child healthcare in low- and middle-income countries. PLOS Medicine. 2017;14(12):e1002465.
325. Opondo C, Allen E, Todd J, English M. Association of the Paediatric Admission Quality of Care score with mortality in Kenyan hospitals: a validation study. The Lancet Global Health. 2018;6(2):e203-e10.
326. Martin S, Duke T, Davis P. Efficacy and safety of bubble CPAP in neonatal care in low and middle income countries: a systematic review. Archives of Disease in Childhood - Fetal and Neonatal Edition. 2014.
327. American Academy of Pediatrics Committee on Fetus and Newborn & American College of Obstetricians and Gynecologists Committee on Obstetric Practice. The Apgar Score. Pediatrics. 2015;136(4):819-22.
328. Wallace E, Smith SM, Perera-Salazar R, Vaucher P, McCowan C, Collins G, et al. Framework for the impact analysis and implementation of Clinical Prediction Rules (CPRs). BMC Medical Informatics and Decision Making. 2011;11(1):62.
329. Harrison W, Goodman D. Epidemiologic trends in neonatal intensive care, 2007-2012. JAMA Pediatrics. 2015;169(9):855-62.
330. Abdel-Latif ME, Nowak G, Bajuk B, Glass K, Harley D. Variation in hospital mortality in an Australian neonatal intensive care unit network. Archives of Disease in Childhood - Fetal and Neonatal Edition. 2017.
331. Sedgwick P. Units of sampling, observation, and analysis. BMJ : British Medical Journal. 2015;351.
332. Mohammed MA, Deeks JJ, Girling A, Rudge G, Carmalt M, Stevens AJ, et al. Evidence of methodological bias in hospital standardised mortality ratios: retrospective database study of English hospitals. Bmj. 2009;338:b780.
333. Lilford R, Pronovost P. Using hospital mortality rates to judge hospital performance: a bad idea that just won't go away. Bmj. 2010;340:c2016.
334. English M, Karumbi J, Maina M, Aluvaala J, Gupta A, Zwarenstein M, et al. The need for pragmatic clinical trials in low and middle income settings – taking essential neonatal interventions delivered as part of inpatient care as an illustrative example. BMC Medicine. 2016;14(1):5.
335. Dickson KE, Simen-Kapeu A, Kinney MV, Huicho L, Vesel L, Lackritz E. Lancet Every Newborn Study Group. Every Newborn: health-systems bottlenecks and strategies to accelerate scale-up in countries. Lancet. 2014;384.
336. Andermann A, Pang T, Newton JN, Davis A, Panisset U. Evidence for Health III: Making evidence-informed decisions that integrate values and context. Health Res Policy Syst. 2016;14:16.
337. Harrison WN, Wasserman JR, Goodman DC. Regional Variation in Neonatal Intensive Care Admissions and the Relationship to Bed Supply. The Journal of Pediatrics. 2018;192:73-9.e4.
338. Rechel B, Wright S, Barlow J, McKee M. Hospital capacity planning: from measuring stocks to modelling flows. Bull World Health Organ. 2010;88(8):632-6.
339. Van Calster B, Van Hoorde K, Vergouwe Y, Bobdiwala S, Condous G, Kirk E, et al. Validation and updating of risk models based on multinomial logistic regression. Diagnostic and Prognostic Research. 2017;1(1):2.
340. Lawn JE, Blencowe H, Darmstadt GL, Bhutta ZA. Beyond newborn survival: the world you are born into determines your risk of disability-free survival. Pediatr Res. 2013;74(S1):1-3.

341. Zeitlin J, Manktelow BN, Piedvache A, Cuttini M, Boyle E, van Heijst A, et al. Use of evidence based practices to improve survival without severe morbidity for very preterm infants: results from the EPICE population based cohort. BMJ. 2016;354.
342. Blencowe H, Lee AC, Cousens S, Bahalim A, Narwal R, Zhong N. Preterm birth-associated neurodevelopmental impairment estimates at regional and global levels for 2010. Pediatr Res. 2013;74.
343. Seale AC, Blencowe H, Zaidi A, Ganatra H, Syed S, Engmann C, et al. Neonatal severe bacterial infection impairment estimates in South Asia, sub-Saharan Africa, and Latin America for 2010. Pediatr Res. 2013;74(S1):73-85.
344. Lee ACC, Kozuki N, Blencowe H, Vos T, Bahalim A, Darmstadt GL, et al. Intrapartum-related neonatal encephalopathy incidence and impairment at regional and global levels for 2010 with trends from 1990. Pediatr Res. 2013;74(S1):50-72.
345. American College of Obstetricians and Gynecologists. Committee Opinion No. 700: Methods for Estimating the Due Date. Obstetrics & Gynecology. 2017;129(5):967-8.
346. Lee AC, Kozuki N, Cousens S, Stevens GA, Blencowe H, Silveira MF, et al. Estimates of burden and consequences of infants born small for gestational age in low and middle income countries with INTERGROWTH-21<sup>st</sup> standard: analysis of CHERG datasets. BMJ. 2017;358.
347. Lassi ZS, Majeed A, Rashid S, Yakoob MY, Bhutta ZA. The interconnections between maternal and newborn health – evidence and implications for policy. Journal of Maternal-Fetal and Neonatal Medicine. 2013;26(S1):3-53.
348. Murphy G, Nyakangi V, Gathara D, Ogero M, English M, null n. A hidden burden of neonatal illness? A cross-sectional study of all admissions aged less than one month across twelve Kenyan County hospitals [version 1; referees: 2 approved]2017.
349. Moons KGM, Altman DG, Reitsma JB, Ioannidis JP, Macaskill P, Steyerberg EW, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): Explanation and Elaboration. Ann Int Med. 2015;162.
350. Altman DG. Prognostic Models: A Methodological Framework and Review of Models for Breast Cancer. Cancer Investigation. 2009;27(3):235-43.
351. Morgan MC, Maina B, Waiyego M, Mutinda C, Aluvaala J, Maina M, et al. Oxygen saturation ranges for healthy newborns within 24 hours at 1800 m. Archives of Disease in Childhood - Fetal and Neonatal Edition. 2017.
352. Ting-Li S, Thomas J, Graeme LH, Iain B, Matthew S. A review of statistical updating methods for clinical prediction models. Statistical Methods in Medical Research. 2016;27(1):185-97.
353. Hippisley-Cox J, Coupland C, Brindle P. Development and validation of QRISK3 risk prediction algorithms to estimate future risk of cardiovascular disease: prospective cohort study. BMJ. 2017;357.

## Appendix

**Appendix 1**: CHARMS 2014 Relevant items to extract from individual studies in a systematic review of prediction models

| Domain | Key items | Reported on page # |
|---|---|---|
| **SOURCE OF DATA** | Source of data (e.g., cohort, case-control, randomized trial participants, or registry data) | |
| **PARTICIPANTS** | Participant eligibility and recruitment method (e.g., consecutive participants, location, number of centers, setting, inclusion and exclusion criteria) | |
| | Participant description | |
| | Details of treatments received, if relevant | |
| | Study dates | |
| **OUTCOME(S) TO BE PREDICTED** | Definition and method for measurement of outcome | |
| | Was the same outcome definition (and method for measurement) used in all patients? | |
| | Type of outcome (e.g., single or combined endpoints) | |
| | Was the outcome assessed without knowledge of the candidate predictors (i.e., blinded)? | |
| | Were candidate predictors part of the outcome (e.g., in panel or consensus diagnosis)? | |
| | Time of outcome occurrence or summary of duration of follow-up | |
| **CANDIDATE PREDICTORS (OR INDEX TESTS)** | Number and type of predictors (e.g., demographics, patient history, physical examination, additional testing, disease characteristics) | |
| | Definition and method for measurement of candidate predictors | |
| | Timing of predictor measurement (e.g., at patient presentation, at diagnosis, at treatment initiation) | |
| | Were predictors assessed blinded for outcome, and for each other (if relevant)? | |
| | Handling of predictors in the modelling (e.g., continuous, linear, non-linear transformations or categorised) | |
| **SAMPLE SIZE** | Number of participants and number of outcomes/events | |
| | Number of outcomes/events in relation to the number of candidate predictors (Events Per Variable) | |
| **MISSING DATA** | Number of participants with any missing value (include predictors and outcomes) | |
| | Number of participants with missing data for each predictor | |
| | Handling of missing data (e.g., complete-case analysis, imputation, or other methods) | |
| **MODEL DEVELOPMENT** | Modelling method (e.g., logistic, survival, neural network, or machine learning techniques) | |
| | Modelling assumptions satisfied | |
| | Method for selection of predictors **for inclusion** in multivariable modelling (e.g., all candidate predictors, pre-selection based on unadjusted association with the outcome) | |
| | Method for selection of predictors **during multivariable modelling** (e.g., full model approach, backward or forward selection) and criteria used (e.g., p-value, Akaike Information Criterion) | |
| | Shrinkage of predictor weights or regression coefficients (e.g., no shrinkage, uniform shrinkage, penalized estimation) | |
| **MODEL PERFORMANCE** | Calibration (calibration plot, calibration slope, Hosmer-Lemeshow test) and Discrimination<br>(C-statistic, D-statistic, log-rank) measures with confidence intervals | |
| | Classification measures (e.g., sensitivity, specificity, predictive values, net reclassification improvement) and whether a-priori cut points were used | |

| | | |
|---|---|---|
| **MODEL EVALUATION** | Method used for testing model performance: development dataset only (random split of data, resampling methods e.g. bootstrap or cross-validation, none) or separate external validation (e.g. temporal, geographical, different setting, different investigators) | |
| | In case of poor validation, whether model was adjusted or updated (e.g., intercept recalibrated, predictor effects adjusted, or new predictors added) | |
| **RESULTS** | Final and other multivariable models (e.g., basic, extended, simplified) presented, including predictor weights or regression coefficients, intercept, baseline survival, model performance measures (with standard errors or confidence intervals) | |
| | Any alternative presentation of the final prediction models, e.g., sum score, nomogram, score chart, predictions for specific risk subgroups with performance | |
| | Comparison of the distribution of predictors (including missing data) for development and validation datasets | |
| **INTERPRETATION AND DISCUSSION** | Interpretation of presented models (confirmatory, i.e., model useful for practice versus exploratory, i.e., more research needed) | |
| | Comparison with other studies, discussion of generalizability, strengths and limitations. | |

**Appendix 2:** CHARMS data extraction template

| Domain | Key items | Reported on page # |
|---|---|---|
| SOURCE OF DATA | | |
| PARTICIPANTS | | |
| | | |
| | | |
| | | |
| OUTCOME(S) TO BE PREDICTED | | |
| | | |
| | | |
| | | |
| | | |
| | | |
| CANDIDATE PREDICTORS (OR INDEX TESTS) | | |
| | | |
| | | |
| | | |
| | | |
| SAMPLE SIZE | | |
| | | |
| MISSING DATA | | |
| | | |
| | | |
| MODEL DEVELOPMENT | | |
| | | |
| | | |
| | | |
| | | |
| MODEL PERFORMANCE | | |
| | | |
| MODEL EVALUATION | | |
| | | |
| RESULTS | | |
| | | |
| | | |
| INTERPRETATION AND DISCUSSION | | |
| | | |

**Appendix 3:** GRADE for Diagnostics

| Category | GRADE for Diagnostics |
|---|---|
| **Question formulation** | PICO formulation; but maybe inadequate due to lack of distinction in type of "patients" and stage of "disease", both elements may have impact on judgments on evidence quality<br><br>Tests need to be considered in the context of test-treatment pathway; development of clinical pathway |
| **Outcomes** | Patient important outcomes<br><br>- Measured directly (randomizing patients to different test-treatment strategies; measurement of patient important outcomes); use GRADE for Interventions approach<br><br>or, in of absence of such direct evidence:<br><br>- Measured indirectly by using diagnostic test accuracy (DTA) outcomes (true positives, false positives, true negatives, false negatives) as surrogate markers for patient important outcomes. Evidence synthesis of DTA data and linkage to patient important outcomes is needed[2]; use GRADE for Diagnostics approach |
| **Rating the quality of evidence** | DTA studies: by outcome, across studies |
| **GRADE criteria for downgrading** | |
| **1. Risk of bias (RoB)** | 3. First, assess RoB for each study. Use RoB domains of the QUADAS 2 tool<br><br>4. Downgrading is based on considered judgment on the extent of bias in all studies reporting the outcome |
| **Category** | **GRADE for Diagnostics** |
| **2. Indirectness** | Downgrading based on applicability issues. E.g. differences between populations studied (e.g. secondary care) and those for whom recommendation or test is intended (e.g. primary care), differences in tests studied or diagnostic expertise of people applying them. Use Applicability domains of QUADAS 2 tool.<br><br>When two or more tests are each compared to its reference standard but not directly to one another:<br>Evidence quality maybe downgraded for indirect comparisons. E.g. (A vs reference test and B vs reference test instead of A vs B directly).<br><br>If focus is on patient important outcomes: quality of evidence downgraded because DTA evidence is considered a surrogate marker for patient important outcomes. |
| **3. Inconsistency** | Downgrading based on evaluation of unexplained heterogeneity.<br><br>Criteria are less clear compared to intervention studies, but include judgment on similarity of point estimates and extent of overlap of confidence intervals. Lack of appropriate statistical methods for assessing study heterogeneity. |
| **4. Imprecision** | Downgrading is based on evaluation of 95% confidence intervals around sensitivity and specificity. Formal guidance and methods on how this can be assessed not available yet. One method might be calculating the projected number of patients testing as FN and FP based on a defined prevalence of the target condition. |
| **5. Publication bias** | Judgment of this criterion is challenging.<br><br>Existing methods used in intervention studies not applicable to diagnostic studies |
| **Category** | **GRADE for Diagnostics** |
| **GRADE criteria for upgrading** | Specific examples of situations in which upgrading is appropriate not yet available. |

[1] Linkage of DTA evidence to patient important outcomes not covered in this table

From: Gopalakrishna G, Mustafa RA, Davenport C, Scholten RJPM, Hyde C, Brozek J, et al. Applying Grading of Recommendations Assessment, Development and Evaluation GRADE) to diagnostic tests was challenging but doable. Journal of Clinical Epidemiology. 2014;67(7):760-8

**Appendix 4:**List of Variables

**1.FULL DATA SET**

- 1086 variables

**2.MINIMUM DATA SET**

- 22 variables

Document Source (NAR)
Date of birth
Admission Date
Today's date
Age (dys)
Gestation (wks)
Residence
Referral
Gender
Birth weight
Weight at admission
Born before arrival?
Mode of delivery
Maternal HIV status
    Maternal Antiretroviral Therapy(ARV/ART/HAART)
    Baby given Antiretroviral Therapy(ARV/ART/HAART)
Admission Diagnoses
Is there a Discharge/Death Summary?
Discharge/ Death Date
Outcome of discharge
Condition on discharge
Discharge Diagnosis
BCG administered during hospitalization
Oral Polio vaccine administered during hospitalization

**Appendix 5**: Ethical review

**Jalemba Aluvaala**

**From:** OxTREC Ethics <oxtrec@admin.ox.ac.uk>
**Sent:** Wednesday, August 31, 2016 3:45 PM
**To:** Jalemba Aluvaala
**Subject:** RE: Clarification on Ethical Review

Dear Jalemba

Many thanks for your email.

I can confirm that, on the understanding that your study uses only previously collected, anonymised data that cann traced back to identifiable individuals, it is exempt from ethical review by OxTREC.

When it comes to future publications, if the journal requires a formal letter (rather than just an email) from OxTREC confirming exemption from ethical review, please let us know and we will be able to provide this for you.

With best wishes

Rebecca

**Dr Rebecca Bryant**
Research Ethics Manager (OxTREC) | Research Ethics & Integrity Team
Research Services, University of Oxford
University Offices, Wellington Square
Oxford OX1 2JD
**T:** +44 (0)1865 282585 **E:** rebecca.bryant@admin.ox.ac.uk
www.admin.ox.ac.uk/researchsupport

# KENYA MEDICAL RESEARCH INSTITUTE

P.O. Box 54840-00200, NAIROBI, Kenya
Tel (254) (020) 2722541, 2713349, 0722-205901, 0733-400003; Fax: (254) (020) 2720030
E-mail: director@kemri.org info@kemri.org Website:www.kemri.org

**KEMRI/RES/7/3/1** **June 16, 2014**

**TO:** **PROF. MIKE ENGLISH**
**PRINCIPAL INVESTIGATOR**

**THROUGH:** **DR. CHARLES MBOGO,**
**ACTING DIRECTOR, CGMR-C,**
**KILIFI**

DIRECTOR CENTRE FOR GLOBAL HEALTH MEDICINE RESEARCH, COAST forwarded 18/6/14

Dear Sir,

**RE: SSC PROTOCOL NO. 2771 (*RESUBMISSION*): A TECHNICAL COLLABORATION WITH THE MINISTRY OF HEALTH TO SUPPORT AND IMPROVE HEALTH SERVICE EVALUATION AND HOSPITAL PERFORMANCE MONITORING- (*VERSION 4 DATED 9TH JUNE, 2014*)**

Reference is made to your letter dated 9th June, 2014. The ERC Secretariat acknowledges receipt of the revised protocol on 12 June 2014.

This is to inform you that the Ethics Review Committee (ERC) reviewed the documents submitted and is satisfied that the issues raised at the 227th meeting of the KEMRI ERC on 20th June, 2014 have been adequately addressed.

Consequently the study is granted approval for implementation effective this **16th June, 2014**. Please note that authorization to conduct this study will automatically expire on **June 15, 2015.** If you plan to continue with data collection or analysis beyond this date, please submit an application for continuing approval to the ERC Secretariat by **May 4, 2015**.

Any unanticipated problems resulting from the implementation of this protocol should be brought to the attention of the ERC. You are also required to submit any proposed changes to this protocol to the SSC and ERC prior to initiation and advise the ERC when the study is completed or discontinued.

You may embark on the study.

Yours faithfully,

EAB

**DR. ELIZABETH BUKUSI,**
**ACTING SECRETARY,**
**KEMRI/ETHICS REVIEW COMMITTEE**

KEMRI / WELLCOME TRUST RESEARCH PROGRAMME
18 JUN 2014
RECEIVED
P. O. Box 230-80108 KILIFI

# KENYA MEDICAL RESEARCH INSTITUTE

P.O. Box 54840 - 00200 NAIROBI - Kenya
Tel: (254) (020) 2722541, 254 (020) 2713349, 0722-205901, 0733-400003 Fax (254) (020) 2720030
Email: director@kemri.org info@kemri.org Website:www.kemri.org

**KEMRI/RES/7/3/1** **May 11, 2016**

DIRECTOR
CENTRE FOR GEOGRAPHIC MEDICINE RESEARCH, COAST
18/5/16

**TO:** **DR. MIKE ENGLISH**
**PRINCIPAL INVESTIGATOR**

**THROUGH:** **DR. BENJAMIN TSOFA**
**THE DIRECTOR, CGMR-C**
**KILIFI**

Dear Sir,

**RE:** **SSC NO: 2771 (*REQUEST FOR ANNUAL RENEWAL)* A TECHNICAL COLLABORATION WITH THE MINISTRY OF HEALTH TO SUPPORT AND IMPROVE HEALTH SERVICE EVALUATION AND HOSPITAL PERFORMANCE MONITORING**

Thank you for the continuing review report for the period **21st April 2015** to **6th April 2016**

This is to inform that during the 251st Committee A meeting of the KEMRI Scientific and Ethics Review Committee held on **10th May, 2016**, the Committee **conducted the annual review and approved** the above referenced application for another year.

This approval is valid from **24th May, 2016** through to **23rd May, 2017**. Please note that authorization to conduct this study will automatically expire on **23rd May, 2017**. If you plan to continue with data collection or analysis beyond this date please submit an application for continuing approval to **SERU** by **11th April 2017.**

You are required to submit any amendments to this protocol and other information pertinent to human participation in this study to SERU for review prior to initiation.

Yours faithfully,

for:

**DR. EVANS AMUKOYE,**
**ACTING HEAD**
**KEMRI/SCIENTIFIC AND ETHICS REVIEW UNIT**

RECEIVED
17 MAY 2016
DIRECTOR'S OFFICE

**Appendix 6:** *List of predictors for selected neonatal prognostic models*

| | | |
|---|---|---|
| **CRIB** | **SNAP** | **NTISS** |
| Birth weight | Blood pressure | Supplemental oxygen |
| Gestation | Heart rate | Surfactant administration |
| Congenital malformation | Respiratory rate | Tracheostomy care |
| Maximum base deficit in first 12 h | Temperature | Tracheostomy placement |
| Minimum appropriate $FIO_2$ in first 12 h | $PO_2$ | CPAP administration |
| Maximum appropriate $FIO_2$ in first 12 h | $PO_2/FIO_2$ ratio | Endotracheal intubation |
| **CRIB II** | $PCO_2$ | Mechanical ventilation |
| Birth weight by gestation | Oxygenation index | Mechanical ventilation with paralysis |
| Maximum base deficit in first 12 h | Packed cell volume | High frequency ventilation |
| Sex | White blood cell count | Extracorporeal membrane oxygenation |
| Admission temperature | Immature total ratio | Indomethacin administration |
| **Berlin score** | Absolute neutrophil count | Volume expansion |
| Birth weight | Platelet count | Vasopressor administration |
| Grade of RDS | Blood urea nitrogen | Pacemaker on standby |
| Apgar score at 5 min | Creatinine | Pacemaker used |
| Artificial ventilation | Urine output | Cardiopulmonary resuscitation |
| Base excess at admission | Indirect bilirubin | Antibiotics |
| **NICHHD score** | Direct bilirubin | Diuretics (enteral) |
| Birth weight | Sodium | Steroids (postnatal) |
| Small for gestational age | Potassium | Anticonvulsant |
| Race | Calcium (ionised) | Aminophylline |
| Sex | Calcium (total) | Other unscheduled medication |
| Apgar score at 1 min | Glucose | Diuretics (parenteral) |
| **NMPI** | Serum bicarbonate | Treatment of metabolic acidosis |
| Gestational age | Serum pH | Potassium binding resin |
| Birth weight | Seizure | Frequent vital signs |
| Cardiac arrest | Apnoea | Cardiorespiratory monitoring |
| $PaO_2/FIO_2$ ratio | Stool guaiac | Phlebotomy |
| Major congenital malformations | **SNAP-PE** | Thermoregulated environment |
| Sepsis | SNAP score plus : | Non-invasive oxygen monitoring |
| Base excess | Birth weight | Arterial pressure monitoring |
| **SINKIN 12 hour** | Apgar score<7 at 5 min | CVP monitoring |
| Birth weight | Small for gestational age | Urinary catheter |
| Gestational age | **SNAP-II** | Quantitative intake and output |
| Apgar score at 5 min | Mean blood pressure | Gavage feeding |
| Peak inspiratory pressure at 12 h | Lowest temperature | Intravenous fat emulsion |
| **NBRS** | $PO_2/FIO_2$ ratio | Intravenous amino acid solution |
| Blood pH | Serum pH | Phototherapy |
| Hypoglycaemia | Multiple seizures | Insulin administration |
| Intraventricular haemorrhage | Urine output | Potassium infusion |
| Periventricular leucomalacia | **SNAPPE-II** | Transfusion |
| Seizures | SNAP II score plus : | Intravenous γ globulin |
| Infection | Birth weight ⩽749 g | Red blood cell transfusion |
| Need for mechanical ventilation | Apgar<7 at 5 min | Partial volume exchange transfusion |
| | Small for gestational age | Platelet transfusion |
| | | White blood cell transfusion |
| | | Double blood cell transfusion |
| | | Transport of patient |
| | | Chest tube |
| | | Minor operation |
| | | Thoracentesis |
| | | Major operation |
| | | Pericardiocentesis |
| | | Pericardial tube |
| | | Dialysis |
| | | Vascular access |
| | | Peripheral intravenous line |
| | | Arterial line |
| | | Central venous line |

CRIB, clinical risk index for babies; $FIO_2$, fractional inspired concentration of oxygen; NBRS, neurobiological risk score; NMPI, neonatal mortality prognosis index; NTISS, neonatal therapeutic intervention scoring system; $PO_2$, partial pressure of oxygen; RDS, respiratory distress syndrome; SNAP, score for neonatal acute physiology; SNAP-PE, score for neonatal acute physiology-perinatal extension.

Reproduced from: Dorling JS, Field DJ. Value and validity of neonatal disease severity scoring systems. Archives of Disease in Childhood - Fetal and Neonatal Edition. 2008;93(2): F80-F2.

**TRIPS & TRIPS II**

Temperature, Respiratory status, Systolic blood pressure, Response to noxious stimuli

**Appendix 7**: Complete search strategy for neonatal treatment intensity scores

**1.PubMed**
**1.1 Primary search**
(((((("Neonatology"[Mesh]) OR "Infant, Newborn"[Mesh])) AND (("Therapeutics"[Mesh]) OR "therapy"[Subheading])) AND intensity) AND scor*.

**1.2 Supplementary searches**
a) Application of the Inui search string
(((((((("Neonatology"[Mesh]) OR "Infant, Newborn"[Mesh])) AND (("Therapeutics"[Mesh]) OR "therapy"[Subheading])) AND intensity))) AND ((Validat* OR Predict* OR Rule*) OR (Predict* AND (Outcome* OR Risk* OR Model*)) OR ((History OR Variable* OR Criteria OR Scor* OR Characteristic* OR Finding* OR Factor*) AND (Predict* OR Model* OR Decision* OR Identif* OR Prognos*)) OR (Decision* AND (Model* OR Clinical* OR Logistic Models/)) OR (Prognostic AND (History OR Variable* OR Criteria OR Scor* OR Characteristic* OR Finding* OR Factor* OR Model*)))

b) Related citations in PubMed

- Related citations for: Neonatal therapeutic intervention scoring system: a therapy-based severity-of-illness index
- Forward citations for: Neonatal therapeutic intervention scoring system: a therapy-based severity-of-illness index

**2.Google scholar**
Forward citations for: Neonatal therapeutic intervention scoring system: a therapy-based severity-of-illness index

**3.EMBASE**
**3.1 Primary search**
((neonat*[Mesh] OR *newborn[Mesh]) AND(treatment[Mesh] OR therap*[Mesh]) AND intensity[Mesh] AND scor*[Mesh]))

**3.2 Supplementary search**
a) Including alternative terms for scor*

((neonat*[Mesh] OR *newborn[Mesh]) AND(treatment[Mesh] OR therap*[Mesh]) AND intensity[Mesh] AND (scor*[Mesh] OR prediction/ OR rule.mp OR computer model/ OR mathematical model/ OR statistical model/ OR stochastic model/ OR model*[Mesh]))

b) Related citations

- Related citations for: Neonatal therapeutic intervention scoring system: a therapy-based severity-of-illness index
- Forward citations for: Neonatal therapeutic intervention scoring system: a therapy-based severity-of-illness index

**4.LILACS (Global Health Library)**
(neonat* OR newborn) AND (therap* OR treat*) AND (scor*)

**5.WPRIM (Global Health Library)**
(neonat* OR newborn) AND (therap* OR treat*) AND (scor*)

**6. IMSEAR (Global Health Library)**
(neonat* OR newborn) AND (therap* OR treat*) AND (scor*)

**7. IMEMR (Global Health Library)**
(neonat* OR newborn) AND (therap* OR treat*) AND (scor*)
**8.CINAHL**
((“neonates or infant or newborn”) AND (treatment or therapy or intervention) AND score)

**9.Bibliographies**
Bibliographies of eligible articles were read to identify any other potentially eligible articles

**Appendix 8**: Search strategy for PubMed for hospital outcomes for neonates in Kenya and Africa

**1.Kenya**

a) Primary search

(((neonat*) OR newborn) AND hospital) AND Kenya

Restricted to title or abstract

Read title then abstract to identify relevant articles

**2.Africa**

a) Primary search

(neonat* OR newborn OR low birthweight OR preterm) AND (hospital OR facilit* OR inpatient) AND (developing OR low income) AND countr*

- Restricted to title or abstract
- Read title then abstract to identify relevant articles

b) Supplementary searches

- ("Infant, Newborn"[Mesh]) AND (("Hospitals"[Mesh]) AND "Africa"[Mesh])
- (morbidity OR mortality OR outcome) AND hospital AND (neonat* OR newborn)
- neonat* OR newborn OR preterm OR low birthweight) AND hospital
- (neonat* OR newborn OR low birthweight OR preterm) AND (hospital care) AND (developing OR low income) AND countr*
- (neonat* OR newborn) AND (morbidity OR mortality) AND hospital

**Appendix 9: R Code**

# Statistical Analyses in R

Jalemba Aluvaala

January 28, 2018

```
#1.LOADING THE REQUIRED PACKAGES
#1.1 Select the packages
  library(pacman)
    pkgs <- c("ProjectTemplate","Hmisc","ggplot2","GGally","pastecs","P
ropCIs",
              "VennDiagram","survival","survminer","cmprsk","epiDisplay
","rms",
              "ROCR","VIM","mice","plotly","gridExtra","xtable","knitr"
,
              "DescTools","plyr","dplyr","CalibrationCurves")

#1.2 Install the packages
   install.packages(pkgs)
   p_load(char = pkgs)

#2.LOADING THE CLEAN DATASETS
  #2.1 This full data set contains 9115 observations
       load("H:/4.DPhil Analysis/DPhil Analysis Aluvaala/cache/cleandat
a.RData")
  #2.2 This data set is for NETS model development (Chapter 5): 5427 ob
servations
      load("H:/4.DPhil Analysis/DPhil Analysis Aluvaala/cache/moddata.R
Data")
  #2.3 This data set is for NETS external validation: 1437 observations
      load("H:/4.DPhil Analysis/Raw Ext Validation Data/valdata.Rda")
  #2.4 SENSS model development
      #a) Load the 31 imputed data sets
          load("H:/4.DPhil Analysis/Imputed dataset/senssdata.Rda")
      #b) Load the  data set used in the imputation model
          load("H:/4.DPhil Analysis/Imputed dataset/missdata3.Rda")
  #2.5 SENSS external validation
     #a)Clean data set
        load("H:/4.DPhil Analysis/Raw Ext Validation Data/cleanvdata.Rd
a")
     #b)Multiply imputed data sets
        load("H:/4.DPhil Analysis/Imputed dataset/senssval.Rda")
     #c)Data set with predictors and auxiliary variables used for imput
ation
        load("H:/4.DPhil Analysis/Imputed dataset/misenss.Rda")
     #d)Load Imputed datasets
  #3.Clinica Utility
    #a)NETS external validation data set and model formula
      load("H:/4.DPhil Analysis/Raw Ext Validation Data/netsdata.Rda")
      load("fit9.RData")# updated model
   #b)SENSS external validation data set
```

```
      load("H:/4.DPhil Analysis/Raw Ext Validation Data/sendata.Rda")
      load("senss2.RData")#updated model

#CHAPTER 4: CASE FATALITY, LENGTH OF STAY AND TIME TO IN-HOSPITAL DEATH
OR DISCHARGE

#1.Description of the data sets
   # Two data sets were used
      #a) Data set 1 (n=9115),named cleandata
          # i)Minimum data set variables collected for all records(9115
)
          # ii)Full data set variables only collected on 60% sample(542
7)

   #1.1 Data set 1
    #a) Numerical
       dim(cleandata)
    #b)Pattern of admissions
      #2014
          plotdata <- subset(cleandata,(cleandata$yradm.factor %in% "20
14"))
          all_adm <- ggplot(plotdata[!is.na(plotdata$madm.factor),],
          aes(fill = plotdata$yradm.factor[!is.na(plotdata$madm.factor)
],
                x = madm.factor))

          p2 <- all_adm + geom_bar(fill = "steelblue") +
            scale_fill_discrete(name = "Year") + ggtitle("2014") +
            theme(plot.title = element_text(hjust = 0.5)) +
            xlab("Month of Admission") +
             theme(axis.title.x = element_text(face = "bold",size = 18,
                                               color = "black")) +
             theme(axis.text.x = element_text(face = "bold",size = 15,
                                              color = "black")) +
            ylab("Number admitted")+
             theme(axis.title.y = element_text(face = "bold",size = 18,
                                               color = "black")) +
             theme(axis.text.y = element_text(face = "bold",
                                              size = 15,color = "black"
))
     #2015
          plotdata2 <- subset(cleandata,(cleandata$yradm.factor %in% "20
15"))
          all_adm <- ggplot(plotdata2[!is.na(plotdata2$madm.factor),],
          aes(fill = plotdata2$yradm.factor[!is.na(plotdata2$madm.factor
)],
              x = madm.factor))

          p3 <- all_adm + geom_bar(fill = "orangered") +
            scale_fill_discrete(name = "Year") + ggtitle("2015") +
            theme(plot.title = element_text(hjust = 0.5)) +
             xlab("Month of Admission") +
              theme(axis.title.x = element_text(face = "bold",
                                               size = 18,color = "black
```

```
")) +
            theme(axis.text.x = element_text(face = "bold",
                                             size = 15,color = "black"
)) +
           ylab("Number admitted") +
            theme(axis.title.y = element_text(face = "bold",size = 18,
                                              color = "black")) +
            theme(axis.text.y = element_text(face = "bold",size = 15,
                                             color = "black")) +
            theme(axis.title.y=element_blank())

  #2.Description of the study population
      lapply(xtics,describe)

  #3.Survival analysis
     #3.1 Case fatality ratio
      #a) Overall
         describe(cleandata$status.factor)
      #b) By sex
        describe(cleandata$status.factor[cleandata$gender.factor %in%
"Male" &
                                           (!is.na(cleandata$los.numer
ic))])
        describe(cleandata$status.factor[cleandata$gender.factor %in%
"Female"&
                                          (!is.na(cleandata$los.numeri
c))])
       #c) By birth weight category
        #Standard categories
        describe(cleandata$status.factor[cleandata$bwtcat %in% "ELBW(<
1)" &
                                           (!is.na(cleandata$los.numer
ic))])
        describe(cleandata$status.factor[cleandata$bwtcat %in% "VLBW(1
-<1.5)" &
                                           (!is.na(cleandata$los.numer
ic))])
        describe(cleandata$status.factor[cleandata$bwtcat %in% "LBW(1.
5-<2.5)"&
                                           (!is.na(cleandata$los.numer
ic))])
        describe(cleandata$status.factor[cleandata$bwtcat %in% "Normal
(2.5-4)"&
                                         (!is.na(cleandata$los.numeric)
)])
        describe(cleandata$status.factor[cleandata$bwtcat %in% "Macros
omia(>4)"&
                                           (!is.na(cleandata$los.numeric)
)])
       #Split LBW(1.5-<2.5)
        describe(cleandata$status.factor[cleandata$bwt4 %in% "1.5-<2"&
                                          (!is.na(cleandata$los.numeric
))])
        describe(cleandata$status.factor[cleandata$bwt4 %in% "2-<2.5"&
```

```
                                        (!is.na(cleandata$los.numeric)
)])
     #3.2 Length of stay (LoS)
      #a)All
        #Numerical
        describe(cleandata$los.numeric)# 8 missing calculated length of
stay

        #Histogram (demonstrates right skewed distribution of LoS)
        lh <- ggplot(cleandata,aes(los.numeric))
        lh + geom_histogram(fill = "steelblue")+
        ggtitle("Density plot for length of stay")
         theme(plot.title = element_text(hjust = 0.5)) +
         theme(plot.title = element_text(face = "bold",size = 20,
                                         color = "black"))+
        xlab("Length of stay (days)")+
         theme(axis.title.x = element_text(face = "bold",
                                           size = 24,color = "black"))+
         theme(axis.text.x = element_text(face = "bold",
                                         size = 20,color = "black"))+
         scale_x_continuous(breaks = seq(0, 100, 10),
                            limits = c(0, 100))+
         theme(axis.title.x = element_text(margin = margin(25,0,0,0)))+
        ylab("Number of patients") +
         theme(axis.title.y = element_text(face = "bold",
                                           size = 24,color = "black"))
+
         theme(axis.text.y = element_text(face = "bold",
                                         size = 20,color = "black"))+
         theme(axis.title.y = element_text(margin = margin(0,25,0,0)))
+
         scale_y_continuous(breaks = seq(0, 4000, 500), limits = c(0, 4
000))

        #Boxplot(Demonstrates outliers and variation by birth weight ca
tegory)
          #All

         ld <- ggplot(cleandata[!is.na(cleandata$bwtcat),],
             aes(x = bwtcat,y = los.numeric,fill = bwtcat))
         ld + geom_boxplot() + ggtitle("Length of stay by birthweight")
+
           theme(plot.title = element_text(hjust = 0.5)) +
           theme(plot.title = element_text(face = "bold",size = 20,
                                           color = "black")) +
         xlab("Birth weight categories (kg)") +
           theme(axis.title.x = element_text(face = "bold",size = 18,co
lor = "black")) +
           theme(axis.text.x = element_text(face = "bold",size = 10,col
or = "black")) +
           theme(axis.title.x = element_text(margin = margin(25,0,0,0))
) +
        ylab("Length of stay (days)") +
```

```
          theme(axis.title.y = element_text(face = "bold",size = 18,col
or = "black")) +
          theme(axis.text.y = element_text(face = "bold",size = 15,colo
r = "black")) +
          scale_y_continuous(breaks = seq(0, 100, 10), limits = c(0,100
)) +
          theme(axis.title.y = element_text(margin = margin(0,25,0,0)))
+
          scale_fill_discrete(name = "Birthweight\nCategory")+
          geom_text(aes(label=..count..), y=100, stat='count', colour="
red", size=4)
         coord_fixed(ratio = 100)

        #Survived to discharge

        ld2 <- ggplot(alivedata[!is.na(alivedata$bwtcat),],
                 aes(x = bwtcat,y = los.numeric,fill = bwtcat))
        ld2 + geom_boxplot() +
         ggtitle("Length of stay by birthweight(Discharged Alive)") +
         theme(plot.title = element_text(hjust = 0.5)) +
         theme(plot.title = element_text(face = "bold",size = 20,
                                         color = "black")) +
       xlab("Birth weight categories (kg)") +
        theme(axis.title.x = element_text(face = "bold",size = 18,
                                          color = "black")) +
        theme(axis.text.x = element_text(face = "bold",size = 10,
                                         color = "black")) +
        theme(axis.title.x = element_text(margin = margin(25,0,0,0))) +
      ylab("Length of stay (days)") +
       theme(axis.title.y = element_text(face = "bold",
                                         size = 18,color = "black")) +
       theme(axis.text.y = element_text(face = "bold",
                                        size = 15,color = "black")) +
       scale_y_continuous(breaks = seq(0, 100, 10),
                          limits = c(0,100)) +
       theme(axis.title.y = element_text(margin = margin(0,25,0,0))) +
       scale_fill_discrete(name = "Birthweight\nCategory")+
       geom_text(aes(label=..count..), y=100, stat='count', colour="red
", size=4)
       coord_fixed(ratio = 100)



    #b) LoS by clinical characteristics
      #Sex
      describe(cleandata$los.numeric[cleandata$gender.factor %in% "Male
"])
      describe(cleandata$los.numeric[cleandata$gender.factor %in% "Fema
le"])

     #Birth weight
      describe(cleandata$los.numeric[cleandata$bwtcat %in% "ELBW(<1)"])
      describe(cleandata$los.numeric[cleandata$bwtcat %in% "VLBW(1-<1.5
```

```
)"])
      describe(cleandata$los.numeric[cleandata$bwtcat %in% "LBW(1.5-<2.
5)"])
      describe(cleandata$los.numeric[cleandata$bwtcat %in% "Normal(2.5-
4)"])
      describe(cleandata$los.numeric[cleandata$bwtcat %in% "Macrosomia(
>4)"])
      describe(cleandata$los.numeric[cleandata$birth_weight < 1.5])
       psych::describe(cleandata$los.numeric[cleandata$birth_weight < 1
.5])
       psych::describe(cleandata$los.numeric[cleandata$birth_weight < 2
])
       psych::describe(cleandata$los.numeric[cleandata$birth_weight >4]
)

     #Outcome
      describe(cleandata$los.numeric[cleandata$otcme.factor %in% "Dead"
])
      describe(cleandata$los.numeric[cleandata$otcme.factor %in% "Alive
"])

   #3.3 Kaplan-Meier Survival

      #a) Survival time in days
        cleandata$time <- cleandata$los.numeric
        cleandata$time[cleandata$otcme.factor%in%"Referred"]<-NA#Referr
ed recoded to missing
        describe(cleandata$time)
      #b) Outcome status
        cleandata$status <- cleandata$otcme.factor
        cleandata$status[cleandata$status %in% "Referred"] <- "Alive" #
since it will be censored recode referred to alive
        cleandata$status <- as.numeric(cleandata$status)
        cleandata$status[cleandata$otcme.factor %in% "Dead"] <- 1
        cleandata$status[cleandata$otcme.factor %in% "Alive" | cleandat
a$otcme.factor %in% "Referred"] <- 0

      #c) Survival object
        cleandata$survobj <- with(cleandata, Surv(time,status))

      #d)Overall K-M
        #K-M Estimator
         km1 <- survfit(survobj~1, data = cleandata)
        #Plot
         ggsurvplot(km1, data = cleandata, risk.table = TRUE,palette  =
"blue",
                           legend = c("none"),
                           legend.labs = c("All"),
                           font.legend =c(17,"bold"),
                           font.x = c(17, "bold"),
                           font.y = c(17, "bold"),
                           font.tickslab = c(12, "bold","black"),
                           fontsize = c(6),
```

```
                          risk.table.fontsize = c(4),
                          ggtheme = theme_minimal())

      #e)K-M by sex
        # K-M Estimator
         km2 <- survfit(survobj~gender.factor,data = cleandata)
        #Plot
         ggsurvplot(km2, data = cleandata, risk.table = TRUE,palette =
"palette",
                    conf.int = FALSE,
                    legend = c("none"),
                    legend.labs = c("Female","Male"),
                    font.x = c(17, "bold"),
                    font.y = c(17, "bold"),
                    font.tickslab = c(12, "bold","black"),
                    fontsize = c(4),
                    ggtheme = theme_minimal())

      #f)K-M by birthweight
        #K-M Estimator
         km3 <- survfit(survobj~bwtcat,data=cleandata)
        #Plot
         ggsurvplot(km3, data = cleandata, risk.table = TRUE,
                    palette = "palette",
                    legend = c("none"),
                    legend.labs = c("ELBW(<1kg)","VLBW(1-<1.5kg)","LBW(
1.5-<2.5kg)"
                                    ,"Normal(2.5-4kg)","Macrosomia(>4kg
)"),
                    font.x = c(17, "bold"),
                    font.y = c(17, "bold"),
                    font.tickslab = c(12, "bold","black"),
                    fontsize = c(4),
                    #risk.table.axis.title.y = c("bold","black"),
                    ggtheme = theme_minimal())

  #3.4 Competing risks survival analysis
    #3.4.1.Required Variables
      #a) Create cleandata$event.type as the event type indicator:
           # 1: Death.
           # 2: Discharged alive.
           # 0: Censored observation: Referred.

           cleandata$event.type <- as.numeric(cleandata$otcme.factor)
           cleandata$event.type[cleandata$otcme.factor %in% "Dead"] <-
1
           cleandata$event.type[cleandata$otcme.factor %in% "Alive"] <-
2
           cleandata$event.type[(cleandata$otcme.factor %in% "Referred"
)
                        & (cleandata$event.type == 3)] <- 0
           cleandata$death <- ifelse(cleandata$event.type == 1,1,0)
      #b) Create variable denoting occurrence of in-hospital death.('1
```

```
1')
        cleandata$dis.alive <- ifelse(cleandata$event.type == 2,1,0)
      #c) Create variable denoting occurrence of discharged alive.('1,2
')
        cleandata$event <- ifelse(cleandata$event.type < 0,1,0)
      #d) Create event indicator for any type of event.

    #3.4.2  Overall competing risk
      #a) Compute CIF
         y <- cuminc(cleandata$time,cleandata$event.type,cencode = 0)
         #41/9115 cases omitted due to missingness (8 time+33 event)
      #b) Plot
        p <- par(mai = c(6, 5, 4, 2)*.25)
         plot(y,
          curvlab = c("Dead", "Alive"),color = 2:1,cex = 1.5,wh = c(70,
0.4),
          xlab = "Time after admission (days)",
          ylab = "Probability",
          col.lab = "black", cex.lab = 2,lwd = 2, line = 4,
          axes = F)
          title("Overall probability of death /discharge alive(n=9074)"
)
          axis(side = 1, at = seq(0.0, 100, by = 14),cex.axis = 1.5)
          axis(side = 2, at = seq(0, 1, by = 0.2),cex.axis = 1.5)
          grid(nx = NULL, ny = NULL, col = "lightgray", lty = "dotted",
          lwd = par("lwd"), equilogs = TRUE)
       par(p)


    #3.4.3 By sex category
      #a) Compute
        x <- cuminc(cleandata$time,cleandata$event.type,
             cencode = 0,group = cleandata$gender.factor)
         #82/9115 cases omitted due to missing values(41 above+41gender
missing)

      #b)Plot
        p <- par(mai = c(6, 5, 4, 2)*.25)
          plot(x,
          curvlab = c("Female-Dead", "Male-Dead","Female-Aive","Male-Al
ive"),color = 1:4,
          wh = c(60,0.42),cex = 1.5,
          xlab = "Time after admission (days)",ylab = "Probability",
          col.lab = "black", cex.lab = 2,lwd = 2,lty = 1, line = 4,
          axes = F)
          title("Probability of death/discharge alive by sex")
          axis(side = 1, at = seq(0.0, 100, by = 14),cex.axis = 1.5)
          axis(side = 2, at = seq(0, 1, by = 0.2),cex.axis = 1.5)
          grid(nx = NULL, ny = NULL, col = "lightgray", lty = "dotted",
          lwd = par("lwd"), equilogs = TRUE)
          par(p)

    #3.4.4 By birthweight category(5 categories)
```

```
    #a)ELBW
       #Compute
       z <- cuminc(cleandata$time,cleandata$event.type,
                   cencode = 0,subset = (cleandata$bwtcat == "ELBW(<1)"
))
             #23 cases omitted due to missingness(from bwtcat overall)
             # None missing for ELBW
      #Plot
       plot(z,
            curvlab = c("Death", "Discharge"),color = 2:1,cex = 1,wh =
c(20,0.6),
            xlab = "Time after admission (days)",ylab = "Probability",
lty = 1,lwd = 2,
             col.lab = "black", cex.lab = 2, axes = F)
            title("Weight<1kg (n=49)")
            axis(side = 1, at = seq(0, 42, by = 7),cex.axis = 1.5)
            axis(side = 2, at = seq(0, 1, by = 0.2),cex.axis = 1.5)
           grid(nx = NULL, ny = NULL, col = "lightgray", lty = "dotted
",
            lwd = par("lwd"), equilogs = TRUE)

    #b)VLBW
      #Compute
       a <- cuminc(cleandata$time,cleandata$event.type,cencode = 0,
                   subset = (cleandata$bwtcat == "VLBW(1-<1.5)"))
                   # 23 cases omitted due to missingness(from bwtcat ove
rall)
                   # None missing for VLBW

       plot(a,
            curvlab = c("Death", "Discharge"),color = 2:1,cex = 1,wh =
c(60,0.6),
            xlab = "Time after admission (days)",ylab = "",lty = 1,lwd
= 2,
             col.lab = "black", cex.lab = 2, axes = F)
          title("Weight 1-<1.5kg (n=280)")
          axis(side = 1, at = seq(0, 98, by = 14),cex.axis = 1.5)
          axis(side = 2, at = seq(0, 1, by = 0.2),cex.axis = 1.5)
          grid(nx = NULL, ny = NULL, col = "lightgray", lty = "dotted"
,
            lwd = par("lwd"), equilogs = TRUE)

       par(q)
    #c)LBW
     #Compute
       b <- cuminc(cleandata$time,cleandata$event.type,cencode = 0,
             subset = (cleandata$bwtcat == "LBW(1.5-<2.5)"))
             #37 cases omitted due to missing values(23 bwtcat overall
+2 time +12 event)
             #14/2132 omitted for LBW
     #Plot
       plot(b,
            curvlab = c("Death", "Discharge"),color = 2:1,cex = 2,wh =
```

```
c(50,0.4),
             xlab = "Time after admission (days)",ylab = "Probability",
lty = 1,lwd = 2,
             col.lab = "black", cex.lab = 2, axes = F)
          #title("Weight 1.5-<2.5kg (n=2118)")
           axis(side = 1, at = seq(0.0, 100, by = 14),cex.axis = 1.5)
           axis(side = 2, at = seq(0, 1, by = 0.2),cex.axis = 1.5)
           grid(nx = NULL, ny = NULL, col = "lightgray", lty = "dotted"
,
             lwd = par("lwd"), equilogs = TRUE)

    #d)Normal Weight
      #Compute
       c <- cuminc(cleandata$time,cleandata$event.type,cencode = 0,
             subset=(cleandata$bwtcat =="Normal(2.5-4)"))
             #48 cases omitted due to missingness (23+5+20)
             #25/6258 missing for Normal weight
      #Plot
       plot(c,
           curvlab=c("Death", "Discharge"),color = 2:1,cex = 1,wh=c(40,
0.4),
           xlab="Survival time (days)",ylab ="Probability",lty = 1,lwd
= 2,
           col.lab = "black", cex.lab = 2, axes = F)
       title("Weight 2.5-4kg (n=6233)")
       axis(side = 1, at = seq(0, 98, by = 14),cex.axis = 1.5)
       axis(side = 2, at = seq(0, 1, by = 0.2),cex.axis = 1.5)
       grid(nx = NULL, ny = NULL, col = "lightgray", lty = "dotted",
           lwd = par("lwd"), equilogs = TRUE)

    #e)Macrosomia
      #Compute
      d <- cuminc(cleandata$time,cleandata$event.type,cencode = 0,
                subset=(cleandata$bwtcat == "Macrosomia(>4)"))
                #24 cases omitted due to missingness(23+0+1)
                #1/353 missing for macrosomia
      #Plot
      plot(d,
           curvlab =c("Death", "Discharge"),color = 2:1,cex = 1,wh =c(2
0,0.4),
           xlab = "Survival time (days)",ylab = "",lty = 1,lwd = 2,
           col.lab = "black", cex.lab = 2, axes = F)
      title("Weight >4kg (n=352)")
      axis(side = 1, at = seq(0, 35, by = 7),cex.axis = 1.5)
      axis(side = 2, at = seq(0, 1, by = 0.2),cex.axis = 1.5)
      grid(nx = NULL, ny = NULL, col = "lightgray", lty = "dotted",
           lwd = par("lwd"), equilogs = TRUE)


  #f)Investigating 1.5-<2.5kg in 500gm increments
    #Compute Wt 1.5-<2kg
       g <- cuminc(cleandata$time,cleandata$event.type,cencode = 0,
                subset = (cleandata$bwt4 == "1.5-<2"))
```

```
                #26 cases omitted due to missingness(23+3)
                #3 missing 1.5-<2
    #Plot Wt 1.5-<2kg
        plot(g,
           curvlab = c("Death", "Discharge"),col = 2:1,cex = 1.5,wh = c
(40,0.4),
           xlab = "Time after admission (days)",ylab = "Cumulative inci
dence",lty = 1,lwd = 2,
           col.lab = "black", cex.lab = 2, axes = F)
          title("Weight 1.5-<2kg (n=670)",cex.main=1.5)
           axis(side = 1, at = seq(0, 98, by = 14),cex.axis = 1.5)
           axis(side = 2, at = seq(0, 1, by = 0.2),cex.axis = 1.5)
          grid(nx = NULL, ny = NULL, col = "lightgray", lty = "dotted",
           lwd = par("lwd"), equilogs = TRUE)

    #Compute Wt 2-<2.5kg
      h <- cuminc(cleandata$time,cleandata$event.type,cencode = 0,
                subset =(cleandata$bwt4 == "2-<2.5"))
                #34 cases omitted due to missingness(23+11)
                #11 missing 2-<2.5kg

    #Plot Wt 2-<2.5kg
      plot(h,
           curvlab = c("Death", "Discharge"),col = 2:1,cex = 1.5,wh = c
(20,0.4),
           xlab = "Time after admission (days)",ylab = "",lty = 1,lwd =
2,
           col.lab = "black", cex.lab = 2, axes = F)
       title("Weight 2-<2.5kg (n=1448)",cex.main = 1.5)
       axis(side = 1, at = seq(0, 42, by = 7),cex.axis = 1.5)
       axis(side = 2, at = seq(0, 1, by = 0.2),cex.axis = 1.5)
       grid(nx = NULL, ny = NULL, col = "lightgray", lty = "dotted",
           lwd = par("lwd"), equilogs = TRUE)

      par(q)



  #3.5 Compare K-M Cumulative incidence against Competing risk
        plot(y,
           curvlab = c("Dead (C.Risks)", "Discharged Alive (C.Risks)"),
           color = 2:1,cex = 1.5,wh = c(60,0.3),
           xlab = "Time after admission (days)",
           ylab = "Cumulative incidence",
           col.lab = "black", cex.lab = 2,lwd = 2, line = 4,
           axes = F)
           axis(side = 1, at = seq(0.0, 100, by = 14),cex.axis = 1.5)
           axis(side = 2, at = seq(0.0, 1, by = 0.1),cex.axis = 1.5)
           grid(nx = NULL, ny = NULL, col = "lightgray", lty = "dotted"
,
           lwd = par("lwd"), equilogs = TRUE)

      par(new = T)
```

```
        plot(km1,fun = "event",col = c("blue"),ylim = c(0,1),conf.int
= F,
          lwd = 1.5,axes = F)
          legend(60,0.35,c("Dead (1-KMeier)"),bty = "n",lty = c(1),lwd
= c(2.5),
            cex = 1.5,col = c("blue"))
      par(p)

# CHAPTER 5:DEVELOPMENT AND VALIDATION OF THE NEONATAL ESSENTIAL TREATM
ENT SCORE (NETS)

  #1. Distribution of essential treatments(predictors) and by outcome
        desct <- moddata[,c("ivadm","ivfadm","birth_weight","bwtcat","f
edadm","oxyadm","phenobarb_all","otcme.factor")]
        lapply(desct, describe)
        sapply(desct,describe)
        tapply(desct$otcme.factor,desct$bwtcat,describe)#split by bwt,d
ecribe outcome to eact bwt category

  #2.Model specification and estimation

    #2.1 Required variables
     #a) First create new outcome variable and recode "referred" to mis
sing
      # Recall that in competitive risk survival analysis referred were
censored
      # Here "Dead"=events "Alive"= non-events
      moddata$status<-moddata$otcme.factor
      moddata$status[moddata$status%in%"Referred"]<-"Alive"
      moddata$status<-as.numeric(moddata$status)
     #b) Create final outcome binary level and re-order levels
      moddata$out.bin<-as.factor(moddata$status)
      levels(moddata$out.bin)[levels(moddata$out.bin)==1]<-0#non-events
      levels(moddata$out.bin)[levels(moddata$out.bin)==2]<-1#events
     #c) Check
      table(moddata$out.bin,moddata$otcme.factor)# appropriately coded
      describe(moddata$out.bin)
     #d) Set normal weight as reference category
      moddata<- within(moddata, bwtcat <- relevel(bwtcat, ref ="Normal(
2.5-4)" ))
      moddata<- within(moddata, bwt4 <- relevel(bwt4, ref = "2.5-4" ))

    #2.2 Model derivation
      fit8 <- lrm(out.bin~bwtcat+gender.factor+ivadm+ivfadm+
                    fedadm+oxyadm+phenobarb_all,data=moddata,
                  x=T,y=T,linear.predictors = T)

    #2.3 Model apparent performance
      #a) Discrimination
         fit8 # print results showing c-statistic:0.918
           # Confidence interval for c-statistic for fit 8
             cstat8 <- rcorr.cens(fit8$linear.predictors, fit8$y)
```

```
              cstat8# gives the c statistic and its SD
              cat(cstat8[1], "[", cstat8[1]-1.96/2*cstat8[3], " - ",
                  cstat8[1]+1.96/2*cstat8[3],"]")
      #b)Calibration curve
          #Required objects
            fit8$p<-1/(1+exp(-fit8$linear.predictors))#linear predictor
(logit,predicted),
            fit8$yn<-as.numeric(levels(fit8$y)[fit8$y])

          #Plot
            cal.plot2<-val.prob.ci.2(fit8$p,fit8$yn,CL.smooth="fill",
                                     logistic.cal=TRUE, lty.log=9,g=10
,
                                     col.log="red", lwd.log=1.5, col.i
deal=colors()[26],
                                     lwd.ideal=2,cex=1.5,cex.leg = 1,c
ex.axis=1.2,
                                     cex.lab=1.3,cex.main=5,font.lab=2
,
                                     #dostats = F)
                           xlab="Predicted Risk",
                           ylab="Observed Deaths")
                           #title("NETS Development, n=4800")

    #2.4 Internal validation
        #a)Bootstrapping
           set.seed(10)
            val.fit8<-validate(fit8,B=1000)#run bootstrap validation fu
nction
            print(val.fit8)
              #Results
                # Discrimination:C-statistic=0.9146548 (see below)
                # Calibration
                   # Intercept:0.0625
                   # Slope :0.9785
        #b) Computation of optimism corrected c-statistic
            cstatv8<-val.fit8[1,1:5]/2 + .5 #(calculated from Somers' D
',select row 1 column 5 of matrix)
            cstatv8 # corrected c-statistic= 0.9157600
        #c) Shrinkage of coefficients to examine for optimism
            #Shrinkage factor
            res.val4 <- val.fit8
            shrinkage.factor4 <- res.val4["Slope","index.corrected"]#sh
rinkage factor
            print(shrinkage.factor4)# Shrinkage factor=0.979,
                                    #this is equivalent to calibration slop
e in Internal Validation
            #Shrunk coefficients
            res.coef5 <- data.frame(Original = coef(fit8),
                                    Shrunk.boot = c(coef(fit8) * shrink
age.factor4))
            round(res.coef5, 3)
```

```
    #3.External validation

      #3.1 Load models
        load("fit8.RData")#Model development
        load("val.fit8.RData")#Internal validation

      #3.2 Required variables
        #a) First create new outcome variable and recode referred to mi
ssing
           # Recall that in competitive risk survival analysis referred
were censored
           valdata$status<-valdata$otcme.factor
           valdata$status[valdata$status%in%"Referred"]<-"Alive"
           valdata$status<-as.numeric(valdata$status)
        #b) Create final outcome binary level and re-order levels
          valdata$out.bin<-as.factor(valdata$status)
          levels(valdata$out.bin)[levels(valdata$out.bin)==1]<-0
          levels(valdata$out.bin)[levels(valdata$out.bin)==2]<-1
        #c) Set normal weight as reference
          valdata<- within(valdata, bwtcat <- relevel(bwtcat, ref ="Nor
mal(2.5-4)" ))

      #3.3 External validation
        #a) Predict (apply development coefficients to validation data
set)
          valdata$p<-predict(fit8,valdata )

        #b)logit of predicted
         valdata$pl<-1/(1+exp(-valdata$p))

        #c)Observed outcome as 1,0
         valdata$obs<-as.numeric(levels(valdata$out.bin)[valdata$out.bi
n])

        #d)Data set without missing
         myvars <- c("survey_id","pl","p","obs","out.bin")#select the v
ariables of interest
         valdata2 <- valdata[myvars]# create new data set
           valdata2 <- subset(valdata,select= c(survey_id,pl,p,obs,out.
bin))
         valdata3<- na.omit(valdata2)# drop the missing

        #e)Model performance(Plot calibration curve)
         exvplot2<-val.prob.ci.2(valdata3$pl,valdata3$obs,CL.smooth="fi
ll",
                                  logistic.cal=TRUE, lty.log=9,g=10
,
                                 col.log="red", lwd.log=2,lwd.smoot
h =2,
                                 col.ideal=colors()[26],lwd.ideal=2
,cex=1.5,
                                 cex.leg = 1,cex.axis=1.2,
                                 cex.lab=1.3,cex.main=5,font.lab=2)
```

```
                                        dostats = F)
                    title("NETS,External Validation (n=1434)")

      #3.4 Update after external validation
        #a)Define the model in the validation data set
           #note the difference in coefficients
           #objective is to obtain the matrix x which has the predictor
s in form of 1,0
           fitup <- lrm(out.bin~bwtcat+gender.factor+ivadm+ivfadm+
                      fedadm+oxyadm+phenobarb_all,data=valdata,
                      x=T,y=T,linear.predictors = T)
           print(fitup)# 9 observations omitted due to missingness (n=1
434:137 events/1297 nonevents)
           exp(coef(fitup))
           print(fit8)

       #b)Coefficients from development model
          print(fit8)
          fit8.par<-c(-4.1521,5.6836,4.5359,1.4186,-0.2927,-0.3125,1.36
95,1.3256,-1.9135,0.6142,2.5947)

       #c)Calculate linear predictor
         lp1 <- fitup$x %*% fit8.par[-1]+fit8.par[1]

       #d)Compute the recalibrated intercept
         lp2 <- lrm.fit(y=fitup$y,offset =lp1)
         print(lp2)

       #e)Update the development model intercept
         fit9<-fit8
         fit9$coefficients[1]<-fit9$coefficients[1]+lp2$coefficients[1]

    #3.5 Performance of final model
        #a) Predict (apply development coefficients to external validat
ion data set)
          valdata$pnew<-predict(fit9,valdata )

        #b)Predicted probabilities
          valdata$plnew<-1/(1+exp(-valdata$pnew))

        #c)Observed outcome as 1,0 in external validation data set
         valdata$obs<-as.numeric(levels(valdata$out.bin)[valdata$out.bi
n])

       #d)Data set without missing
          myvars <- c("survey_id","plnew","pnew","obs","out.bin")#selec
t the variables of interest
          valdata4 <- valdata[myvars]# create new data set
          valdata5<- na.omit(valdata4)# drop the missing

       #e)Plot calibration curve
        exvplot3<-val.prob.ci.2(valdata5$plnew,valdata5$obs,CL.smooth="
fill",
```

```
                                      logistic.cal=TRUE, lty.log=9,g=10,
                                    col.log="red", lwd.log=2, col.ideal=c
olors()[26],
                                    lwd.ideal=2,lwd.smooth= 2,cex=1.5,cex
.leg = 1,
                                    cex.axis=1.2,
                                    cex.lab=1.3,cex.main=5,font.lab=2,dos
tats = F)

#CHAPTER 6:DEVELOPMENT AND VALIDATION OF THE SCORE OF ESSENTIAL NEONATA
L SYMPTOMS AND SIGNS (SENSS)
   #1.Multiple Imputation
     #1.1 Development data set
       #a) Subset to random sample data which contains symptoms and sig
ns predictors
          impdata<-subset(cleandata,(cleandata$randomized.factor=="No")
)
          dim(impdata)#5427 1086

       #b) Describe distribution of predictors
          desc_senss <- impdata[,c("otcme.factor","gender.factor","bwtc
at",
                                   "difficulty_feeding.factor","all.con
vulsions",
                                   "central_cyanosis.factor","all.indra
wing",
                                   "floppy_nosuck","bulging_fontanelle.
factor")]
          lapply(desc_senss, describe)

      #c) Create data set with predictors, auxilliary variables and out
come
          myvars3 <- c("survey_id","gender.factor","delivery.factor","b
wtcat",
                       "bba.factor","maternal_hiv.factor","apg_doc.fact
or",
                       "difficulty_feeding.factor","all.convulsions",
                       "central_cyanosis.factor","all.indrawing","flopp
y_nosuck",
                       "otcme.factor")
          missdata3 <- impdata[myvars3]

      #d) Missingness pattern
          missdata_aggr = aggr(missdata3, col=c("Skyblue","Yellow","Gre
en"),
                               numbers=TRUE, sortVars=TRUE, labels=name
s(missdata),
                           cex.axis=1,ylim=0.2,
                           gap=3, ylab=c("Proportion of missingness",
                                         "Missingness Pattern"))
      #e) Required variables
         # Recall that in competitive risk survival analysis referred w
ere censored
           missdata3$otcme.factor[missdata3$otcme.factor%in%"Referred"]
```

```
<-"Alive"
          missdata3$otcme.factor<-as.numeric(missdata3$otcme.factor)
        #Create final outcome binary level and re-order levels
          missdata3$otcme.factor<-as.factor(missdata3$otcme.factor)
          levels(missdata3$otcme.factor)[levels(missdata3$otcme.factor
)==1]<-0
          levels(missdata3$otcme.factor)[levels(missdata3$otcme.factor
)==2]<-1
        #Set reference categories
          missdata3<- within(missdata3, bwtcat <- relevel(bwtcat,
                                                          ref ="Normal(2.
5-4)" ))
          missdata3<- within(missdata3,
                             difficulty_feeding.factor<- relevel(difficu
lty_feeding.factor, ref ="No" ))
          missdata3<- within(missdata3,
                             all.convulsions<- relevel(all.convulsions,r
ef ="No" ))
          missdata3<- within(missdata3,
                             all.indrawing <- relevel(all.indrawing, ref
="No" ))
          missdata3<- within(missdata3,
                             central_cyanosis.factor<- relevel(central_c
yanosis.factor, ref ="No" ))
          missdata3<- within(missdata3,
                             floppy_nosuck <- relevel(floppy_nosuck, ref
="No" ))

      #f) Mutliple Imputation
          missdata3$delivery.factor[missdata3$delivery.factor%in%"Assi
sted vaginal (Includes Forceps,Vacuum)"] <- "Spontaneous vaginal (SVD)"
            missdata3<-droplevels.data.frame(missdata3)
          senssdata<-aregImpute(formula=~survey_id+gender.factor+deliv
ery.factor+
                                bwtcat+bba.factor+maternal_hiv.factor
+
                                apg_doc.factor+difficulty_feeding.fac
tor+
                                all.convulsions+central_cyanosis.fact
or+
                                all.indrawing+floppy_nosuck+
                                otcme.factor,data=missdata3,n.impute=
31)

     #1.2 External validation data set
      #a)Subset to random sample data which contains symptoms and signs
predictors
        impdata<-subset(cleanvdata,(cleanvdata$randomized.factor=="No"
))
        dim(impdata)#1627 739

      #b)Data frame with predictors,auxiliary variables and outcome onl
y
```

```
         myvars2 <- c("survey_id","gender.factor","delivery.factor",
                      "bwtcat","bba.factor","maternal_hiv.factor",
                   "apg_doc.factor","difficulty_feeding.factor","all.con
vulsions",
                "central_cyanosis.factor",
                "all.indrawing","floppy_nosuck","otcme.factor")
         missdata <- impdata[myvars2]
      #c)Missingness pattern
         missdata_aggr = aggr(missdata, col = c("Skyblue","Yellow","Gre
en"),
                              numbers = TRUE,
                              sortVars = TRUE, labels = names(missdata),
                     cex.axis =.5,ylim = 0.2,
                     gap = 3, ylab = c("Proportion of missingness",
                                       "Missingness Pattern"))
      #d)Impute
         #Required varaibles

         # Recall that in competitive risk survival analysis referred w
ere censored
            missdata$otcme.factor[missdata$otcme.factor%in%"Referred"]<
-"Alive"
             missdata$otcme.factor<-as.numeric(missdata$otcme.factor)
         #Create final outcome binary level and re-order levels
            missdata$otcme.factor<-as.factor(missdata$otcme.factor)
            levels(missdata$otcme.factor)[levels(missdata$otcme.factor)
==1]<-0
           levels(missdata$otcme.factor)[levels(missdata$otcme.factor)=
=2]<-1
         #Set reference categories
          missdata<- within(missdata, bwtcat <- relevel(bwtcat,
                                                        ref ="Normal(2.
5-4)" ))
          missdata<- within(missdata,
                  difficulty_feeding.factor<- relevel(difficulty_feedin
g.factor,
                                                      ref ="No" ))
          missdata<- within(missdata,
                            all.convulsions<- relevel(all.convulsions,re
f ="No" ))
          missdata<- within(missdata,
                     all.indrawing <- relevel(all.indrawing, ref ="No"
))
          missdata<- within(missdata,
                     central_cyanosis.factor<- relevel(central_cyanosi
s.factor
                                                       , ref ="No" ))
          missdata<- within(missdata,
                            floppy_nosuck <- relevel(floppy_nosuck, ref
="No" ))

          misenss <- missdata
        #Impute
          misenss$delivery.factor[misenss$delivery.factor%in%"Assisted
```

```
vaginal (Includes Forceps,Vacuum)"] <- "Spontaneous vaginal (SVD)"
          misenss<-droplevels.data.frame(misenss)

          senssval<-aregImpute(formula=~survey_id+gender.factor+deliver
y.factor+
                                  bwtcat+bba.factor+maternal_hiv.factor+
                            apg_doc.factor+difficulty_feeding.factor+
                              all.convulsions+central_cyanosis.factor+
                            all.indrawing+floppy_nosuck+otcme.factor,
                            data=misenss,n.impute=23)

  #2.Model specification and estimation
      #2.1 Derivation
         senss2 <- fit.mult.impute(otcme.factor~bwtcat+gender.factor+
                                    difficulty_feeding.factor+all.convuls
ions+
                                     all.indrawing+central_cyanosis.fact
or+
                                    floppy_nosuck,lrm,xtrans=senssdata,
                                    data=missdata3,x=T,y=T,linear.predic
tors = T)
     #2.2 Model apparent performance
        load("senss2.RData")
       #a)Discrimination
        senss2 # C statistic:0.90
        # Confidence interval for c-statistic for fit 8(from Steyerburg
t821.24sept09.r.txt)
        cstat2 <- rcorr.cens(senss2$linear.predictors, senss2$y)
        cstat2# gives the c statistic and its SD
        cat(cstat2[1], "[", cstat2[1]-1.96/2*cstat2[3], " - ", cstat2[1
]+1.96/2*cstat2[3],"]")
        #0.911128 [ 0.8950621  -  0.9271939 ]

       #b) Calibration
         #First create objects as expected in funtion for predicted and
observed
         senss2$p<-1/(1+exp(-senss2$linear.predictors))#predicted proba
bility of the event(0-1)
         senss2$yn<-as.numeric(levels(senss2$y)[senss2$y])# observed(1
or 0 as its binary)

         #Calibration plot
         tiff("H:/4.DPhil Analysis/Thesis Figures/senss_devt.tiff",widt
h = 8,
             height = 8, units = 'in', res = 300)
             senssplot2<-val.prob.ci.2(senss2$p,senss2$yn, CL.smooth="f
ill",
                                       smooth=c("rcs"),CL.BT = F,nr.kno
ts = 3,
                                       logistic.cal=TRUE ,lty.log=9,g=1
0,
                                       col.log="red", lwd.log=2,lwd.smo
oth = 2,
```

```
                                        col.ideal=colors()[26],lwd.ideal
=2,
                                        cex=1.5,cex.leg = 1,
                                        cex.axis=1.2,cex.lab=1.3,cex.mai
n=2,
                                        font.lab=2)# dostats = F)
                                        xlab="Predicted Risk"
                                        ylab="Observed Deaths"
                                        #title("SENSS,Development, n=542
7")

      #2.3 Internal validation
        #a) Bootstrap
          val.senss2<-validate(senss2,B=500)

        #b)Optimism corrected c-statistic
          csenss2<-val.senss2[1,1:5]/2 + .5 #(calculated from Somers' D
',select row 1 column 5 of matrix)
          csenss2 # corrected c-statistic= 0.8985181

        #c)Shrinkage of coefficients
          #Compute shrinkage factor
            shrinksenss2 <- val.senss2["Slope","index.corrected"]
            print(shrinksenss2)#0.9859403

          #Shrunk coefficients
            res.coef3 <- data.frame(Original = coef(senss2),
                                    Shrunk.boot = c(coef(senss2) * shrin
ksenss2))
            round(res.coef3, 3)

   #3.External validation
      #3.1 Development model
        load("senss2.RData")#model

      #3.2 Predict in imputed data sets
        #Prepare
           data = misenss
           xtrans = senssval
           n.impute = xtrans$n.impute#number of imputations
        #Predict
        library(Amelia)
          pred<-NULL
          se.pred<-NULL
              for(i in 1:n.impute) 99

          combined.results <- mi.meld(q = pred, se = se.pred)
          combined.predictions <- as.list(combined.results$q.mi)

      #3.3 Performance
          #a)Prepare predictions
             misenss$y <- as.numeric(levels(misenss$otcme.factor)[misen
ss$otcme.factor])# Convert observed outcome to numeric
```

```
            misenss$p <- as.numeric(combined.predictions)#Let the pred
iction be the combined predictions after multiple imputation
          #b) Validate and plot
            plot3 <- val.prob.ci.2(misenss$p,misenss$y,CL.smooth="fill
",
                                   lwd.smooth = 2,logistic.cal=TRUE,
                                   lty.log=9,g=10,col.log="red", lwd.l
og=2,
                                   col.ideal=colors()[26], lwd.ideal=2
,
                                   cex=1.5,cex.leg = 1,
                                   cex.axis=1.2,cex.lab=1.3,cex.main=2
,
                                   font.lab=2) #dostats = F)
                                   xlab="Predicted Risk"
                                   ylab="Observed Deaths"
                                   title("SENSS,External Validation (n
=1627)")
    #3.4 Update model
        #a) Define the model in the validation data set
            #Obtain the matrix x
            senssup1<- fit.mult.impute(otcme.factor~bwtcat+gender.facto
r+difficulty_feeding.factor+all.convulsions+
                                     all.indrawing+central_cyanosis.facto
r+floppy_nosuck,
                                         lrm,xtrans=senssval,data=misenss
,x=T,y=T,linear.predictors = T)
       #b)Coefficients from development model
        print(senss2)
        senss2.par<-c(-3.8583,5.7580,3.7082,0.9232,-0.4918,-0.1336,1.35
96,1.3977,1.9790,0.9584,1.6266)

       #c)Calculate linear predictor
       lp1 <- senssup1$x%*%senss2.par[-1]+senss2.par[1]

      #d) Compute the recalibration(the intercept only)
        lp2 <- lrm.fit(y=senssup1$y,offset =lp1)
        print(lp2)

      #e) Update the development model intercept
       senssup2<-senss2
       senssup2$coefficients[1]<-senssup2$coefficients[1]+lp2$coefficie
nts[1]

   #3.5 Performance of final model
      #a) Prepare
        data=misenss
        xtrans=senssval# same as in areg.impute
        n.impute = xtrans$n.impute#number of imputations

      #b)Predict
        library(Amelia)
        pred<-NULL
```

```
        se.pred<-NULL
         for(i in 1:n.impute) {
         completed.data <- data
         imputed.data <- impute.transcan(xtrans, imputation = i,
                                         data=data, list.out = TRUE, pr
= FALSE, check = FALSE)
         completed.data[names(imputed.data)] <- imputed.data

        lp <- predict(senssup2,completed.data,se.fit = TRUE)# predictor
log scale,predict function
          p <-1/(1+exp(-lp$linear.predictors))#probability scale
          lp.n<-as.numeric(lp$linear.predictors)# as numeric vector
          p.n<-as.numeric(p)# as numeric vector
          z<-lp$se.fit# standard error
          pred <- rbind(p.n)
         se.pred <- rbind(z)
       }

       combined.update <- mi.meld(q = pred, se = se.pred)
       combined.up_pred <- as.list(combined.update$q.mi)

     #c)Calibration plot
       load("H:/4.DPhil Analysis/Imputed dataset/combined.up_pred.Rda")

        misenss$y<-as.numeric(levels(misenss$otcme.factor)[misenss$otcm
e.factor])# Convert observed outcome to numeric
        misenss$p <- as.numeric(combined.up_pred)#Let the prediction be
the combined predictions after multiple imputation

        plot4 <- val.prob.ci.2(misenss$p,misenss$y,CL.smooth="fill",lwd
.smooth = 2,
                               logistic.cal=TRUE, lty.log=9,g=10,
                              col.log="red", lwd.log=2, col.ideal=color
s()[26],
                              lwd.ideal=2,cex=1.5,cex.leg = 1,cex.axis=
1.2,
                              cex.lab=1.3,cex.main=2,font.lab=2) #dosta
ts = F)
        title("SENSS,external validation with updated intercept (n=1627
)")

#CHAPTER 7:EVALUATION OF THE UTILITY OF THE NETS AND SENSS MODELS IN CL
INICAL DECISION MAKING
  #1.Decision curve analysis
    #1.1 NETS
       #a)Predict
        netsdata$netsmodel <- with(netsdata,predict(fit9,netsdata))
        netsdata$netsmodel <- 1/(1 + exp(-netsdata$netsmodel))
      #b)dca
        set.seed(20)
        netsdca2 <- decision_curve(out.bin~netsmodel,
                                   data = netsdata,
                                   fitted.risk = TRUE,
```

```
                                    thresholds = seq(0, 1, by = .1),
                                    bootstraps = 2000)
    #1.2SENSS
     #a) Predict
       sendata$senmodel <- with(sendata,predict(senss2,sendata))
       sendata$senmodel <- 1/(1 + exp(-sendata$senmodel))
     #b)dca
       set.seed(20)
       sendca2 <- decision_curve(otcme.factor~senmodel,
                                 data = sendata,
                                 fitted.risk = TRUE,
                                 thresholds = seq(0, 1, by = .1),
                                 bootstraps = 2000)
    #1.3 Plot
       p <- par(mai = c(8, 5, 4, 2)*.25)
            plot_decision_curve(sendca2,  curve.names = "SENSS",confid
ence.intervals = F,
                                 xlab = "Theshold probability",ylab = "
Net benefit",
                                 cost.benefit.xlab = "Threshold probabi
lity(Pt):1-Pt",
                                 cex.lab = 1.5,cex.axis=1.5,col = c("bl
ue"))
         par(p)

   #2.Clinical impact
      par(mfrow = c(1, 2))
        #p <- par(mai = c(8, 5, 4, 2)*.25)
        #NETS
        plot_clinical_impact(
          netsdca2,
          xlim = c(0, .8),
          confidence.intervals = F,
          main = "Clinical impact of NETS",
          xlab = "Theshold probability",
          cost.benefit.xlab = "Threshold probability(Pt):1-Pt",
          cex.lab = 1.5,
          cex.axis = 1.5,
          col = c("red", "black"),
          legend.position = "right"
        )

        #SENSS
        plot_clinical_impact(
          sendca2,
          xlim = c(0, .8),
          confidence.intervals = F,
          main = "Clinical impact of SENSS",
          xlab = "Theshold probability",
          ylab = "",
          cost.benefit.xlab = "Threshold probability(Pt):1-Pt",
          cex.lab = 1.5,
          cex.axis = 1.5,
          col = c("blue", "black"),
```

```
    legend.position = "right"
  )
```

**Appendix 10:** Newborn Admission Record (NAR)

| **Baby's details** | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Baby Name** | | | **Date of Admission** | | dd/mm/yyyy | | **IP No.** | | | |
| **DOB** | | **Age** | days | hrs | **Sex** | **F** ☐ | **M** ☐ | **Birth Wt** | kg | |
| **Gestation** | weeks | **Temp (⁰C)** | | | **Apgar** | 1m | 5m | 10m | **Wt now** | kg |
| **BBA?** | **Y** ☐ | **N** ☐ | ***..if Yes born where*** | | ☐ Home | ☐ Clinic | | ☐ Other hospital | | |
| **Delivery** | SVD | Vacuum | Breech | CS | **Resuscitation** | None | Oxygen | Bag/Mask | | |
| ***If CS reason*** | ☐ Elective | ☐ Emergency | | | **Placenta** | | | | | |
| **ROM** | <12h | 12-18h | >18h | **Given Vitamin K** | **Y**☐ | **N**☐ | **Given Eye prophylaxis** | **Y**☐ | **N**☐ | |

| **Mother's details** | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **Name** | | | **Age** | | **IP No.** | | | | |
| **Residence sublocation** | | | **Gravidity** | | **Parity** | | | | |
| **HIV** | ☐ **Pos** ☐ **Neg** | **ARVs** | ☐ **Yes** ☐ **No** | **VDRL** | ☐ **Pos** ☐ **Neg** | **Blood Gp** | | | |
| **Fever** | **Y**☐ | **N**☐ | **Antibiotics** | **Y**☐ | **N**☐ | **Diabetes** | **Y**☐ | **N**☐ | **TB treat** | **Y**☐ | **N**☐ |
| **Labour** | **1st Stg** hr | **2nd Stg** mins | | | | **Hypertension** | **Y**☐ | **N**☐ | **APH** | **Y**☐ | **N**☐ |

| **Babies Presenting Problems / Mothers Problems & relevant drugs pre-admission** |
|---|
| |

| **Vital Signs** | **Temp (⁰C)** | | **Resp Rate** | bpm | **Pulse** | /min | **O2 Sat** | % |
|---|---|---|---|---|---|---|---|---|

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| ***Time baby seen*** | am/pm | | | | Stridor | **Y**☐ | **N**☐ |
| Length of illness | days | | | | Cry | ☐Normal ☐Weak ☐Hoarse | |
| Fever – No. of days = | **Y**☐ | **N**☐ | | | Central Cyanosis | **Y**☐ | **N**☐ |
| Difficulty breathing | **Y**☐ | **N**☐ | | **A & B** | Indrawing | ☐None/mild ☐Severe ☐Sternum | |
| Diarrhoea: No. of days = | **Y**☐ | **N**☐ | | | Grunting | **Y**☐ | **N**☐ |
| Diarrhoea bloody | **Y**☐ | **N**☐ | | | Air entry bilateral | **Y**☐ | **N**☐ |
| Vomits everything | **Y**☐ | **N**☐ | | | Crackles | **Y**☐ | **N**☐ |
| Difficulty feeding | **Y**☐ | **N**☐ | | | Cry | ☐Normal ☐Hoarse ☐Weak | |
| Convulsions: No. = | **Y**☐ | **N**☐ | | **C** | Femoral Pulse | ☐ Normal | ☐ Weak |
| Partial / focal fits? | **Y**☐ | **N**☐ | | | Cap Refill | secs | *X = not possible* |
| Apnoea | **Y**☐ | **N**☐ | | | Murmur **Y**☐ **N**☐ | ***If yes =*** | |

| General Examination | | | |
|---|---|---|---|
| Skin | ☐Bruising ☐Rash ☐Pustules | | |
| Jaundice | ☐None | ☐ + | ☐+++ |
| Gest/Size | ☐Normal ☐Prem ☐SGA/wasted | | |
| ***Abnormalities – tick all relevant & describe***<br>☐Skull =<br>☐Limbs / Spine =<br>☐Palate / Face =<br>☐Genitals / anus =<br>☐Dysmorphic = | | | |

| | | | | |
|---|---|---|---|---|
| **C** | Pallor / Anaemia | **0** | **+** | **+++** |
| | Skin cold: | ☐Hand ☐Elbow ☐Shoulder | | |
| **Disability** | Can suck / breastfeed? | **Y☐** | **N☐** | |
| | Stiff neck | **Y☐** | **N☐** | |
| | Bulging fontanelle | **Y☐** | **N☐** | |
| | Irritable | **Y☐** | **N☐** | |
| | Reduced movement / floppy | **Y☐** | **N☐** | |
| ***Umbilicus*** | ☐Clean ☐Local pus ☐Pus + red skin | | | |

**Summary**

**Investigations ordered – record results in medical record**

| | | | |
|---|---|---|---|
| ***Malaria*** | ☐ Blood slide ☐ Rapid Test | ***Glucose*** | ☐ Stick test ☐ Laboratory |
| ***Haematology*** | ☐ Hb ☐ HCT ☐ Full haemogram | ***Chemistry*** | ☐ Na + K ☐ Urea ☐ Creat ☐ LFT |
| ***Microbiology*** | ☐ Lumbar Punct ☐ Blood Cult | ***HIV*** | ☐ Rapid test ☐ PCR |
| ***X-Ray*** | CXR ☐ AXR ☐ Other = | ***Other 1*** | |

**Admission Diagnoses – Select ONE *primary diagnosis (tick box indicating "1") and ANY secondary diagnoses (tick box indicating "2")***

| | | | | | | |
|---|---|---|---|---|---|---|
| ***Birth asphyxia*** | 1☐ | 2☐ | ***Meconium aspiration*** | 1☐ | 2☐ | ***Maternal HIV Status*** ☐Pos ☐Neg ☐Unknown |
| ***Premature / LBW*** | 1☐ | 2☐ | ***Twin delivery*** | 1☐ | 2☐ | ***Other diagnoses (name below and indicate if primary diagnosis or secondary)*** |
| ***Newborn RDS*** | 1☐ | 2☐ | ***Jaundice*** | 1☐ | 2☐ | 1☐ 2☐ |
| ***Neonatal sepsis*** | 1☐ | 2☐ | ***Meningitis*** | 1☐ | 2☐ | 1☐ 2☐ |

**Supportive care – indicate what care you are providing a plan for and sign please**

| | | | | |
|---|---|---|---|---|
| Vitamin K & TEO | ❑ | ARVs for PMTCT | ❑ | **Clinician Name & Sig** |
| Nutrition / Feeds | ❑ | Oxygen | ❑ | |
| iv fluids | ❑ | Blood transfusion | ❑ | |
| Incubator / Keep warm` | ❑ | Phototherapy | ❑ | |

| **Results of Investigations** | Date / Test / Results | Date / Test / Results | Date / Test / Results |
|---|---|---|---|
| | | | |

**Appendix 11: Search strategy for PubMed for Decision curve analysis applied to neonatal prediction models**

((((((("infant, newborn"[MeSH Terms] OR neonatal[Text Word])) AND ((Validat$ OR Predict$.ti. OR Rule$) OR (Predict$ AND (Outcome$ OR Risk$ OR Model$)) OR ((History OR Variable$ OR Criteria OR Scor$ OR Characteristic$ OR Finding$ OR Factor$) AND (Predict$ OR Model$ OR Decision$ OR Identif$ OR Prognos$)) OR (Decision$ AND (Model$ OR Clinical$ OR Logistic Models/)) OR (Prognostic AND (History OR Variable$ OR Criteria OR Scor$ OR Characteristic$ OR Finding$ OR Factor$ OR Model$))))) AND ((Decision OR decide))) AND analy*) AND curve*