

**DEVELOPMENT OF PROGNOSTIC MODELS IN THE
PRESENCE OF COMPETING RISKS AND MODEL
UNCERTAINTIES: AN APPLICATION TO
PEADIATRIC IN-HOSPITAL MORTALITY**

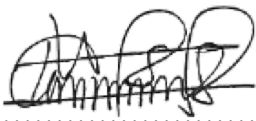
Morris Ogero Ondieki

A thesis submitted to the Department of Mathematics, University of
Nairobi for the award of Degree of Doctor of Philosophy in
Biostatistics.

6 October 2023

Declaration


This thesis is my original work and has not been presented for the award of a degree in any other university.

Signature  Date..... 05-12-2023

Morris Ogero Ondieki

Registration number: I80/53692/2018

This thesis has been submitted to examination with our approval as university supervisors.

Signature  Date..... 06-12-2023

Dr. John Ndiritu

Department of Mathematics, University of Nairobi, Kenya

Signature  Date..... 07-12-2023

Dr. Rachel Sarguta

Department of Mathematics, University of Nairobi, Kenya

Signature  Date..... 07-12-2023

Dr. Samuel Akech

KEMRI-Wellcome Trust Research Programme, Kenya

Abstract

This report delves into the practical applications of advanced modeling in the realm of prognostic modeling for pediatric in-hospital mortality. This research commenced with a comprehensive systematic review aimed at identifying predictive models for in-hospital mortality among pediatric patients in resource-limited settings. While the review unearthed twenty-one prognostic models from fifteen studies, it also unveiled significant methodological concerns. These included issues such as poor reporting, suboptimal handling of missing data, inadequate sample sizes, and misjudged categorization of continuous predictors, which collectively cast doubt on the models' predictive capabilities. Subsequently, the research progressed to external validation, assessing the predictive ability of the identified models using data from pediatric patients in 20 county referral hospitals between 2014 and December 2021. Of the 21 models, only 4 met the criteria for external validation. The validation metrics encompassed discriminatory ability (c-statistics) and model calibration (slope and intercepts). The findings consistently revealed a trend of underestimating the risk of mortality in all four models, highlighting the potential for misclassifying high-risk patients. To rectify the miscalibration issue, the focus shifted towards recalibrating these models. Two recalibration strategies were explored, with logistic recalibration proving more effective. However, the improvements, while notable, did not meet the necessary clinical standards, primarily due to a lack of consideration of model uncertainty during the development of the individual models in their original studies. Addressing the pivotal problem of model uncertainty, a stacking of model predictive distributions methodology was introduced. This innovative approach merged predictive distributions from four distinct models (which were refitted), enhancing the accuracy and reliability of mortality risk predictions. When comparing the performance of the individual models with the stacked posterior distribution, the latter surpassed individual models, offering improved discrimination and calibration, promising significant advancements in predictive

accuracy. The focus then shifted to the Fine-Gray Sub-distribution Hazard model in the context of competing risks, with Monte Carlo simulations revealing the impact of patient follow-up duration on model accuracy. The findings underscored the challenges in managing competing risks and the limitations of established approaches, particularly in epidemiological research. In conclusion, this study embarks on an innovative journey into the development, validation, and recalibration of prognostic models for predicting in-hospital mortality in pediatric patients. It underscores the importance of ensemble techniques in mitigating model uncertainties and improving predictive accuracy. Despite remarkable progress, further research is needed to address the intricacies of competing risks and enhance model reliability.

To my beloved late dad,

Your fervent desire for my academic success echoes profoundly in my heart and drives my pursuit of excellence. Your teachings, wisdom, and unwavering guidance continue to shape my endeavors. Though you're no longer with me, your spirit lives on in the pages of this thesis.

To my dear mum,

Your prayers and boundless love have been my source of strength and inspiration. Your unwavering belief in my abilities has been a guiding light throughout this journey.

To my brother John (Blaze),

Your constant well-wishes and encouragement during my PhD research journey have been a source of immense motivation. Your support and belief in my aspirations mean the world to me.

To my wife Sylvia, Baraka & Bahati

Your unwavering support and understanding have been the cornerstone of my resilience. Your encouragement and patience throughout the compilation of this PhD thesis have been invaluable.

To my cherished family members,

This work is dedicated to each one of you, whose love, support, and encouragement have been the pillars upon which this achievement stands. Thank you for being my unwavering support system.

Acknowledgments

I would like to express my deepest gratitude to my esteemed PhD supervisors, Dr. Sam Akech, Dr. John Ndiritu, and Dr. Rachel Sarguta. Their unwavering support, invaluable guidance, and scholarly insights have been the pillars of this research journey, and I am immensely grateful for their mentorship.

My sincere thanks go to Professor Mike English and the entire Clinical Information Network (CIN) team for generously permitting the use of their invaluable data in this research. Without their support and data provision, this work would not have been possible.

I extend my heartfelt appreciation to Professor Edwine Barasa for his pivotal role in allowing me to register as a PhD student.

I would also like to express my profound appreciation to the Wellcome Trust Research Programme for their generous support in disseminating the research findings. Their financial assistance in covering the publication costs of my manuscripts is deeply appreciated.

Lastly, I wish to convey my appreciation to the Higher Education Loans Board for their scholarship, which provided crucial financial support for covering my tuition fees throughout my academic journey at the University of Nairobi. Their assistance has been vital in turning my academic aspirations into a reality.

List of publications

The contents of this thesis are based on the following manuscripts.

1. **Ogero M**, Ndiritu J, Sarguta R, Tuti T, Akech S (2023). Pediatric prognostic models predicting inhospital child mortality in resource-limited settings: An external validation study. *Health Sciences Reports*.
<https://onlinelibrary.wiley.com/doi/10.1002/hsr2.1433>
2. **Ogero M**, Ndiritu J, Sarguta R, Tuti T, Aluvaala J, Akech S (2023). Recalibrating prognostic models to improve predictions of in-hospital child mortality in resource-limited settings. *Paediatric Perinatal Epidemiology*.
<https://onlinelibrary.wiley.com/doi/10.1111/ppe.12948>
3. **Ogero M**, Sarguta R, Malla L, Aluvaala J, Agweyu A, Akech S (2020). Methodological rigor of prognostic models for predicting in-hospital paediatric mortality in low- and middle-income countries: a systematic review protocol. *Wellcome Open Research*.
<https://wellcomeopenresearch.org/articles/5-106/v1>
4. **Ogero M**, Sarguta RJ, Malla L, Aluvaala J, Agweyu A, English M, Onyango NO, Akech S (2020). Prognostic models for predicting in-hospital paediatric mortality in resource-limited countries: a systematic review. *BMJ Open*
<https://bmjopen.bmj.com/content/10/10/e035045>
5. **Ogero M**, Malla L, Akech S (2020). Examining which clinicians provide admission hospital care in a high mortality setting and their adherence to guidelines: an observational study in 13 hospitals. *BMJ- Archives of Disease in Childhood*
<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7361020/>
6. **Ogero M**, Ndiritu J, Sarguta RJ, Akech S. Assessing methodological utility of sub-distribution hazard model in patients with short follow-up period in the setting of competing risks. *Paediatric Perinatal Epidemiology*
7. **Ogero M**, Ndiritu J, Sarguta RJ, Akech S. Accounting for model uncertainty through stacking of predictive distributions of prognostic models. *Statistics in Medicine*

List of tables

Table 2.2.3-1: Systematic review framework as recommended by CHARMS checklist.	11
Table 2.2.3-2: Search terms for prognostic models	11
Table 4.2.1-1: Models to be externally validated.	47
Table 4.2.4-1: Predictors used in RISC-Malawi model and their level of missingness both in derivation and validation datasets.	53
Table 4.2.4-2: Predictors used in Lowlaavar et al. model and their level of missingness both in derivation and validation datasets.	54
Table 4.3.2-1: Demographic and clinical characteristics of the cohort used to externally validate RISC-Malawi model.	56
Table 4.3.3-1: Demographic and clinical characteristics patients from 6 hospitals who were included in the validation cohort of Lowlaavar et al. 2016 models.	58
Table 4.3.5-1: Demographic and clinical characteristics of patients included in the sensitivity analyses dataset.	62
Table 5.2.5-1: Minimum required sample sizes for recalibration of identified models.	74
Table 5.3.1-1: Distribution of clinical characteristics of the cohort used to recalibrate and test RISC-Malawi model.	78
Table 5.3.1-2: Demographic and clinical characteristics of the cohort used to recalibrate and test Lowlaavar model.	79
Table 5.3.2-1: Correction factors for model intercept and model slope	80
Table 6.2.1-1: Prognostic factors included in the 4 predictive models.	89
Table 6.3.1-1: cohort of patients used to train and test models.	83
Table 6.3.3-1: Individual model performances compared to that of ensemble methods.	97
Table 7.3.1-1: Distribution of clinical characteristics of the empirical sample used for data generation.	111

List of figures

- Figure 2.3.1-1: *PRISMA flow diagram showing the process used to identify prognostic models predicting in-hospital paediatric mortality included in this review.* 15
- Figure 2.3.1-2: *Prognostic models predicting in-hospital paediatric mortality identified by country. Text highlighted in red are the names of the models with their corresponding discrimination measures (area under the curve). Key: PEWS-RL score=Paediatric Early Warning Score for Resource-Limited Settings; SICK score=Signs of Inflammation in Children that Kill; PET score= Paediatric Emergency Triage; mRISC score= Modified Respiratory Index of Severity in Children score; RISC score= Respiratory Index of Severity in Children score; PERCH severity score= Pneumonia Etiology Research for Child Health severity score; LOD score= Lambarene Organ Dysfunction score; CRT= Classification and Regression Trees; ITAT Score= Inpatient Triage Assessment and Treatment score; PEDIA score= Paediatric Early Death Index for Africa score.* 16
- Figure 2.3.2-1: *Top four categories of predictors in the models of the reviewed reports: altered consciousness (coma, prostration, not alert, unconscious); malnutrition indicators (kwashiorkor, edema, weight-for-height z-score, weight-for-age z-score, mid-upper arm circumference-MUAC, wasting); vital signs (temperature, respiratory rate, heart rate, oxygen saturation); signs of respiratory distress (indrawing, lung crepitation, difficult breathing, grunting).* 18
- Figure 2.3.2-2: *Summary of the risk of bias of the included models using PROBAST (Prediction study Risk of Bias Assessment Tool).*..... 21
- Figure 2.3.2-3: *Risk of bias assessment. Low means low risk of bias, High means a high risk of bias, and Unclear bias means it was not possible to assess the risk of bias. Key: PEWS-RL score=Paediatric Early Warning Score for Resource-Limited Settings; SICK score=Signs of Inflammation in Children that Kill; PET score= Paediatric Emergency Triage; mRISC score= Modified Respiratory Index of Severity in Children score; RISC score= Respiratory Index of Severity in Children score; PERCH severity score= Pneumonia Etiology Research for Child Health severity score; LOD score= Lambarene Organ Dysfunction score; CRT= Classification and Regression*

Trees; ITAT Score= Inpatient Triage Assessment and Treatment score; PEDIA score= Paediatric Early Death Index for Africa score. 22

Figure 3.2.1-1: Locations of hospitals included in the validation cohort. 32

Figure 3.3-1: Kernel density plots of the observed (non-imputed) and imputed values of various variables. Visual inspection of the distributions of the observed and imputed values appears identical suggesting the imputation model generated plausible values to replace missing ones. 36

Figure 3.7.1-1: Schematic view of the competing event framework for hospital admissions. Every admitted patient experience any of the four mutually exclusive events (discharged, referred out, discharged against the advice or absconded, and died). Cause-specific hazard functions are denoted by α_0jt , $j = 1, 2, 3, 4$ and each arrow denotes a transition to a state denoted by a rectangle. 43

Figure 4.2.2-1: Patients meeting the eligibility criteria of inclusion for external validation of various models. 50

Figure 4.2.2-2: Locations of the 6 hospitals whose patients were included in the validation cohort of Lowlaavar et al. models. 52

Figure 4.3.4-1: Performance of the RISC-Malawi model in an external validation dataset. The figures show calibration curves and other model performance metrics. Key: RCS denotes the Restricted Cubic Splines, and CL denotes the Confidence Limits (95%)..... 59

Figure 4.3.4-2: Performance of the Lowlaavar et al. 2016 models in an external validation dataset where abnormal Blantyre Coma Score (BCS) was defined as $BCS < 5$. The first panel to the left is the calibration curves of the primary model (Model I), the panel in middle are the calibration curves for model II, and the last panel to the right are the calibration curves of the model III. Key: RCS denotes the Restricted Cubic Splines, and CL denotes the Confidence Limits (95%)..... 60

Figure 4.3.5-1: Performance of the RISC-Malawi model in a sensitivity analyses dataset (Pneumonia is defined based on the admission clinical diagnosis instead of danger signs). The values show calibration curves and other model performance metrics. Key: RCS denotes the Restricted Cubic Splines, and CL denotes the Confidence Limits (95%). 61

Figure 4.3.5-2: Performance of the Lowlaavar et al. 2016 models in an external validation dataset whereby Abnormal Blantyre coma score was defined using the disability scale of AVPU (Alert, Verbal response, Pain response, Unresponsive) such that patients who were not alert but responding to verbal stimuli were assumed to have abnormal Blantyre coma score. The first panel to the left is the calibration curves of the primary model (Model I), the panel in middle are the calibration curves for model II, and the last panel to the right are the calibration curves of the model III. Key: RCS denotes the Restricted Cubic Splines, and CL denotes the Confidence Limits (95%)..... 63

Figure 5.2.2-1: Model intercept and model slope of the four models suggesting that models were not well calibrated. These estimates were obtained from external validation study 71

Figure 5.2.5-1: Populations used to update and test RISC-Malawi model and 3 models by Lowlaavar et al. 2016 75

Figure 5.3.2-1: RISC-Malawi model calibration performance in various datasets. The figure in the left show calibration intercept while that on the right shows model slope. The coloured points and the 95% confidence intervals (shown as errors bars) shows the model calibration performances in the external validation, updating dataset (for model recalibration), and in the testing dataset. The dotted line denotes the references of the model intercept($\alpha=0$) and slope($\beta=1$) for a perfect calibrated model. 81

Figure 5.3.2-2: Calibration performance of Lowlaavar models in various datasets. The figure in the left show calibration intercept while that on the right shows model slope. The coloured points and the 95% confidence intervals (shown as errors bars) shows the model calibration performances in the external validation, updating dataset (for model recalibration), and in the testing dataset. The dotted line denotes the references of the model intercept($\alpha=0$) and slope($\beta=1$) for a perfect calibrated model. 82

Figure 5.3.2-3: Discriminatory ability of the four models (RISC-Malawi, and the 3 models by Lowlaavar et al.) in various datasets. The coloured points and the 95% confidence intervals (shown as errors bars) shows the c-statistics of the in the derivation dataset, external validation, updating (for

model recalibration), and in the testing dataset. The dotted line denotes a fair discriminatory ability of the model (Area Under Curve of 0.7) 83

To find the best possible values of time at which subjects are censored, optimal values of the rate parameters λ_{censor} of the exponential distribution were required. For each scenario and in each iteration a new seed was set that corresponded to the iteration number. Setting of seed ensured the reproducibility (reuse of the same set of random variables) of the sequence where necessary. Exponential random variables were simulated from a huge population ($N=1,000,000$). This was informed from previous simulation studies. Optimal λ_{censor} values were searched in the interval $[0.01, 500]$ to achieve the desirable proportion of patients for whom an event was observed to occur in the simulated data using a bisection algorithm. This approach starts with a large interval known to contain the solution, then it successively reduces interval size until the solution is found. Our stopping criteria was defined as follows; if the difference between the probability of being censored in the empirical dataset and in the simulated dataset was negligible (<0.0001) for any given scenario. The schematic view of how this was implemented is as shown in Figure 7.2.6-1..... 107

Figure 7.2.6-2: The bisection algorithm used to search for the optimal rate parameter for the exponential distribution that was used to simulate time at which patients were censored. 108

Figure 7.3.1-1: Distributions of the patient outcomes across hospitals 110

Figure 7.3.2-1: The y-axis is the relative bias in recovering the true model coefficient for variable severe acute malnutrition at different values of the parameter p (0.1, 0.5, and 0.9) across various sample sizes (500, 1000, and 10000). 112

Figure 7.3.2-2: The y-axis is the relative bias in recovering the true model coefficient for variable child sex at different values of the parameter p (0.1, 0.5, and 0.9) across various sample sizes (500, 1000, and 10000). 113

Figure 7.3.2-3: The y-axis is the relative bias in recovering the true model coefficient at different values of the parameter p (0.1, 0.5, and 0.9) across various sample sizes (500, 1000, and 10000). The upper panel represent the variable temperature while the lower panel represent the variable age in months..... 114

List of Abbreviations

WHO	World Health Organization
LMIC	Low- and Middle-Income Countries
CoxPH	Cox Proportional Hazards
DGM	Data-generating model
PROSPERO	Prospective Register of Systematic Reviews
CHARMS	Checklist for critical Appraisal and data extraction for systematic Reviews of prediction Modelling Studies
AUC	Area Under the Curve
PROBAST	Prediction study Risk of Bias Assessment Tool
EPV	Events per variable
RoB	Risk of Bias
CCA	Complete Case Analysis
MAR	Missing at Random
MCAR	Missing Completely at Random
MICE	Multiple Imputation by Chained Equations
LOESS	Locally estimated scatterplot smoothing
CIN	Clinical Information Network
REDCap	Research Electronic Data Capture
LOS	Length of hospital stay
ETAT	Emergency Triage Assessment and Treatment
RISC	Respiratory Index of Severity in Children
MUAC	Mid-upper arm circumference
AVPU	Alert, Verbal, Pain, Unresponsive
BCS	Blantyre Coma Score

CI	Confidence interval
AIC	Akaike Information Criterion
BIC	Bayesian Information Criterion
MCMC	Monte Carlo Markov Chain
KNN	K-Nearest Neighbors
LGBM	Light Gradient Boosting Machine
LOO	Leave-One-Out cross-validation
SH	Sub-distribution hazard model

Contents

Declaration.....	i
Abstract.....	i
Acknowledgments.....	iv
List of publications	v
List of tables.....	vi
List of figures.....	vii
List of Abbreviations	xi
Chapter 1.....	1
1.0 General Introduction	1
1.1 Background.....	1
1.1.1 Competing events.....	2
1.1.2 Model uncertainty	3
1.2 Statement of the problem	4
1.3 Objectives of the study.....	5
1.3.1 General objective	5
1.3.2 Specific objectives	5
1.4 Significance of the study.....	6
1.5 Thesis outline.....	7
Chapter 2.....	8
2.0 Literature Review.....	8
2.1 Introduction.....	8
2.2 Methods.....	9
2.2.1 Protocol and registration	9
2.2.2 Eligibility criteria	9
2.2.3 Search strategy of articles	10
2.2.4 Data extraction from the included articles	12
2.2.5 Assessment of methodological rigor of the identified prognostic models.....	13
2.3 Results.....	14
2.3.1 Characteristics of the included studies.....	14
2.3.2 Methodological issues of the reviewed models	17
2.4 Discussion.....	23
2.4.1 Summary of key findings.....	23
2.4.2 Comparison with other Studies	25
2.4.3 Implications of this review.....	27

2.4.4	Strengths and limitations of this review.....	28
2.5	Conclusion	29
Chapter 3.....		31
3.0	Research Methodology	31
3.1	Introduction.....	31
3.2	Data.....	31
3.2.1	Data sources.....	31
3.2.2	Data management.....	33
3.2.3	Missing data imputation.....	34
3.3	Prognostic model’s external validations	37
3.3.1	Determining model’s predicted risks in the validation set.....	37
3.3.2	Determining model’s performance in the validation set.....	38
3.4	Model recalibration.....	39
3.4.1	Model recalibration strategies.....	39
3.4.2	Adjusting model intercept only/recalibration-in-the-large.....	40
3.4.3	Logistic calibration	40
3.5	Accounting for model uncertainty	41
3.5.1	Stacking of predictive distributions	41
3.6	Competing risk framework and simulations	41
3.6.1	Overview of competing risks framework.....	42
3.6.2	Patient survival/follow-up period.....	43
Chapter 4.....		44
4.0	External Validation of Prognostic Models.....	44
4.1	Introduction.....	44
4.2	Methods.....	45
4.2.1	Prognostic models included for external validation.....	45
4.2.2	Model validation dataset	49
4.2.3	Assessing performance of the prognostic model in the external validation.....	53
4.2.4	Missing data in the model validation set.....	53
4.3	Results.....	55
4.3.1	Eligible population.....	55
4.3.2	Characteristics of the cohort used in the external validation of the RISC-Malawi prognostic model.....	56
4.3.3	Characteristics of the cohort used in the external validation of the Lowlaavar <i>et al.</i> 2016 models. 57	
4.3.4	Model performances in external validation dataset.....	58
4.3.5	Sensitivity Analyses.....	61

4.4	Discussion.....	64
4.4.1	Summary of key findings.....	64
4.4.2	Limitations.....	66
4.4.3	Fulfilled knowledge gaps and what to be done next.....	66
4.5	Conclusions.....	67
Chapter 5.....		68
5.0	Recalibrating Prognostic Models.....	68
5.1	Introduction.....	68
5.2	Methods.....	69
5.2.1	Models' calibration metrics.....	69
5.2.2	Details of models to be recalibrated.....	69
5.2.3	Availability of model predictors in the recalibration cohort.....	72
5.2.4	Eligibility criteria for model recalibration cohort.....	72
5.2.5	Sample size for model recalibration.....	73
5.2.6	Assessment of missing data in the model recalibration cohort.....	76
5.2.7	Model recalibration strategy.....	76
5.2.8	Assessing performance of the recalibrated prognostic model in the testing dataset.....	77
5.3	Results.....	77
5.3.1	Characteristics of the model recalibrating cohorts.....	77
5.3.2	Predictive performance of the recalibrated RISC-Malawi model.....	79
5.4	Discussion.....	83
5.5	Conclusion.....	86
Chapter 6.....		87
6.0	Accounting for model uncertainty through stacking of predictive distributions of prognostic models.....	87
6.1	Introduction.....	87
6.2	Methods.....	89
6.2.1	Prognostic factors considered in development of the meta model.....	89
6.2.2	Model computations.....	90
6.2.3	Model fitting.....	90
6.2.4	Model averaging methods.....	90
6.2.5	Model performance assessment.....	92
6.3	Results.....	93
6.3.1	Characteristics of the cohort used to train and test models.....	93
6.3.2	Result of model's stacking weights.....	94
6.3.3	Comparison of model performances.....	96

6.4	Discussion.....	98
6.4.1	Summary of the findings.....	98
6.4.2	Stacking of predictive distribution method vs other approaches.	99
6.4.3	Limitations.....	100
6.4.4	Implications of the study findings.....	100
6.5	Conclusion.....	101
Chapter 7.....		102
7.0	Quantifying the Impact of Short Follow-Up Period on the prognostic model: A Monte Carlo Simulation Study.....	102
7.1	Introduction.....	102
7.2	Methods.....	103
7.2.1	Monte Carlo simulations.....	103
7.2.2	Simulation scenarios.....	103
7.2.3	Data-generating process.....	104
7.2.4	Model coefficients based on the empirical data.....	105
7.2.5	Simulation of event types and time-to-event.....	105
7.2.6	Searching rate parameter of exponential distribution through bisection method.....	107
7.2.7	Model coefficients based on the simulated data.....	109
7.2.8	Assessing estimation bias.....	109
7.3	Results.....	109
7.3.1	Characteristics of the empirical sample used for data generation.....	109
7.3.2	Relative bias in the estimated quantities.....	111
7.4	Discussion.....	115
7.4.1	Principal findings.....	115
7.4.2	Strengths of the study.....	116
7.4.3	Limitations of the data.....	117
7.4.4	Conclusions.....	117
Chapter 8.....		118
8.0	Conclusion, Recommendations and Further Research.....	118
References.....		121

Chapter 1

General Introduction

1.1 Background

Over the past decades, there has been a considerable progress in improving care, but child mortality remains high in sub-Saharan Africa relative to the rest of the world[1, 2]. Paediatric deaths in hospitalized children mostly occur soon after admission [3] and are caused by common childhood illnesses such as malaria, pneumonia, etc., which are readily treatable by cost-effective interventions[4]. To reduce child mortality and morbidity in Low-and Middle-Income Countries (LMIC), World Health Organization (WHO) recommend use of a set of clinical signs to identify children whose health status is at the risk of deterioration for immediate treatment[5]. However, due to multifactorial nature of making clinical predictions, clinicians have difficulty to objectively and simultaneously weigh multiple risk factors to produce reliable and accurate predictions. Therefore, there is need for prognostic models that estimate the actual individual risk as accurately as possible.

A prognostic model is a mathematical equation used to quantify the risk that a patient will experience an event of interest (e.g. death) in a specified time-period [6]. Such models are useful tools for clinicians and patients especially in screening for high-risk patients who could benefit from prompt management especially in LMIC where mortality is high. These prognostic models have been increasingly published over the last three decades [7]. However, majority of these prognostic models are yet to gain wide acceptance in clinical practice due to some reasons including models not being externally validated as should be [8] and poor predictive performances when subjected to external validations. Consequently, clinicians have continued to use their cognitive bias and gut feelings to predict patients' possible outcomes [9]

despite evidence from numerous studies which have suggested that gut feelings are frequently wrong on the predictions of mortality[10]. This problem is further compounded by a high patient-clinician ratio in LMIC[11]. A clinically useful prognostic model should not only be derived using appropriate methodology but should also have clinical relevance [12]. In this report we focus on two methodological issues that include failure to account for competing events, and model uncertainty in the development of prognostic models.

1.1.1 Competing events

The conventional approach to analyzing time-to-in-hospital mortality typically relies on the Kaplan-Meier method for estimating survival functions and the Cox Proportional Hazards (CoxPH) model for gauging the impact of covariates on the hazard function, as outlined in prior research[13]. This method adeptly tackles the challenging issue of censoring, where the ultimate survival time remains unknown, by assuming non-informative censoring[14]. This assumption posits that individuals who are censored and those still at risk share similar prospects for survival. Additionally, it presumes that the reasons behind censoring are unrelated to the study's objectives[15]. However, this assumption has its limitations, particularly in scenarios where patients are discharged alive from the hospital upon recovery or when deteriorating patients are referred elsewhere for specialized care. In such cases, these individuals possess distinct characteristics from those still within the hospital, leading to different survival prospects. The presence of competing events is a common occurrence in biomedical research, yet traditional survival analysis methods continue to be employed, despite their tendency to yield biased estimates[16].

To address this limitation, it is imperative to adopt methods that relax these assumptions. The competing risk framework offers a solution by considering alternative outcomes as competing events that preclude the occurrence of the event of interest[17]. For instance, being discharged alive or being referred are deemed competing events, with either one of these outcomes

preventing the in-hospital death (the event of interest) from occurring. Nonetheless, the suitability of the competing-risk framework in studies with short survival times remains uncertain[18]. Therefore, there is a need to explore its effectiveness and applicability in such contexts.

1.1.2 Model uncertainty

For many years, the central concern in the realm of predictive modeling has revolved around mitigating model uncertainty, especially when the primary objective is accurate out-of-sample prediction, which places the utmost value on model predictive performance [19]. In the domain of prognostic modeling, researchers and statisticians commonly employ data-driven or computer-assisted techniques to determine the covariates to include in a specific regression model. This selection process often relies on information criteria such as AIC or BIC. Subsequently, inferences are drawn under the presumption that the chosen "optimal" model serves as the true data-generating model (DGM).

While this conventional approach is prevalent in the literature and adheres to established statistical practices, it falls short in effectively addressing model uncertainty during the covariate selection process. Consequently, this approach frequently leads to an underestimation of the uncertainties associated with the quantity of interest, ultimately yielding unreliable inferences[20, 21].

In this thesis, the aim is to address the following key questions; what is the utility of competing risks framework in the setting of short survival time such as 2 days? What is the prediction accuracy of the existing prognostic models? What modelling techniques to use in handling model uncertainty?

1.2 Statement of the problem

In Low- and Middle-Income Countries (LMIC), there is a concerning trend of pediatric ward deaths occurring within a short time after admission, typically within 48 hours. To address this issue effectively and identify patients at risk of deterioration, it is crucial to employ prognostic models. While numerous such models exist in the literature, none have received a recommendation for clinical use, especially in high-mortality settings. This lack of endorsement stems from several key factors. Firstly, most of these models have not undergone external validation, and for the few that have, their predictive performance has proven subpar. One primary reason behind this poor performance is the use of sub-optimal methodologies, including the failure to consider competing risks and model uncertainty.

Consequently, there are several critical issues that need addressing. First, the clinical utility of these published models remains uncertain because there are limited validation studies evaluating and comparing their performance in larger cohorts of similar patients. This uncertainty means that these models cannot be relied upon for clinical use until their effectiveness is demonstrated. Second, the potential of the competing-risk framework in developing prognostic models in situations where the time-to-event is short, and competing events are significant, has not been explored adequately. Third, considering the high patient-clinician ratio in most LMIC public hospitals, there is an urgent need for a robust prognostic model to predict mortality in children with short survival periods and thus save lives in this setting. Lastly, addressing uncertainty in prognostic models is vital, especially when dealing with an overwhelming number of possible models to consider from a single dataset or when the number of explanatory covariates far exceeds the available data.

This thesis aims to tackle these issues comprehensively with the goal of improving the out-of-sample predictive accuracy of prognostic models in the future.

1.3 Objectives of the study

1.3.1 General objective

The main objective of this study is to advance the field of pediatric predictive modeling by enhancing the methodological quality, clinical utility, and robustness of existing models.

1.3.2 Specific objectives

The specific objectives of this study are to:

1. Conduct a systematic review to assess and rank published pediatric predictive models based on their methodological quality. This evaluation will specifically focus on examining how these models address issues related to model uncertainty and competing events, if at all.
2. Evaluate the clinical usefulness and reliability of the models identified in objective 1 through an external validation study.
3. Perform model recalibration on existing prognostic models to enhance their predictive performance.
4. Develop a robust method for addressing model uncertainty during the development of a predictive model.
5. Investigate the methodological suitability of the sub-distribution hazard model for accounting for competing risks or events in situations characterized by a short time to event.

1.4 Significance of the study

The significance of this thesis lies in its profound implications for the field of pediatric healthcare and predictive modeling. Firstly, the systematic review outlined in Objective 1 is poised to provide a comprehensive understanding of the landscape of pediatric predictive models. Its significance extends to shedding light on both the strengths and limitations of existing models. Knowledge gained from that piece of work is invaluable for clinicians who rely on these tools to make critical decisions about the care and treatment of pediatric patients. Additionally, it offers researchers and practitioners a well-informed starting point for further advancements in the field.

Objective 2, which involves the external validation of identified models, bridges the gap between theoretical efficacy and real-world applicability. The significance here is two-fold: it offers empirical evidence of the utility and reliability of these models, thus guiding clinical practice, and it ensures that predictive models remain relevant and trustworthy in the ever-evolving landscape of pediatric healthcare.

Objective 3, which is focusing on model recalibration, addresses a fundamental concern in predictive modelling -the need for models to adapt and remain accurate over time. The ability to fine-tune these models based on empirical data can significantly enhance their predictive power, ultimately leading to more precise clinical decisions.

Objective 4 takes on the formidable challenge of addressing model uncertainty during model development. The significance of this lies in its potential to bolster the trustworthiness of predictive models. By devising robust methods to manage model uncertainty, this study promotes accuracy and reliability of model predictions.

Lastly, Objective 5 delves into the methodological suitability of the sub-distribution hazard model for handling competing risks in pediatric scenarios marked by short time-to-event windows. This has direct implications for clinical practice, ensuring that prognostic

assessments accurately reflect the complex realities of pediatric healthcare, reducing the risk of inappropriate interventions or resource allocation.

In conclusion, this thesis embodies a multifaceted endeavor with significant implications for the pediatric healthcare landscape. Its objectives collectively contribute to the improvement of predictive modeling techniques, enhance the quality of patient care, and provide a solid foundation for future research in the field.

1.5 Thesis outline

The structure of this report unfolds as follows: In Chapter 2, we embark on a comprehensive literature review of prognostic models. Chapter 3 provides a detailed overview of the research methodology and introduces key concepts. The external validation of prognostic models is the focus of Chapter 4. In Chapter 5, a thorough exploration of model recalibration is conducted to enhance predictive ability. Chapter 6 delves into innovative approaches for addressing model uncertainties. Chapter 7 involves extensive Monte Carlo simulations to investigate the impact of a short follow-up period on the accuracy of prognostic models in the context of competing events. Finally, Chapter 8 draws this report to a close with a discussion and recommendations for areas warranting further research.

Chapter 2

Literature Review

2.1 Introduction

Child survival has improved significantly over the last few decades; however, in sub-Saharan Africa, child mortality remains disproportionately high compared to the global average. Pediatric deaths within hospitals typically occur shortly after admission, often due to treatable conditions like malaria, pneumonia, and diarrheal diseases. These diseases can be managed effectively with cost-efficient interventions [3, 22, 23]. In low- and middle-income countries (LMICs), healthcare practitioners frequently rely on clinical signs, as recommended by the World Health Organization (WHO) guidelines, to identify patients at risk of deterioration and make informed treatment decisions [24]. The WHO's clinical criteria, developed based on expert recommendations and a review of relevant studies reporting mortality risk factors, serve as the basis for these guidelines. To further enhance patient outcomes, prognostic or predictive models, using statistical equations and a combination of risk factors, can assist clinicians in identifying high-risk patients [25].

Despite numerous prognostic models for hospitalized children being published over the last three decades [7]; there are concerns regarding the methodology employed in their development [26]. Notably, none of these models are recommended for use in resource-limited settings according to current clinical practice guidelines. Consequently, there is a need for reviews of the methodologies underpinning their development [27].

This chapter aims to address this gap by identifying and summarizing studies that have developed prognostic models or scoring systems to predict in-hospital pediatric mortality in

LMICs. Specifically, it provides a comprehensive overview of the existing research and critically assesses the methodological robustness of each model.

2.2 Methods

2.2.1 Protocol and registration

Following recommendations, a research protocol for this review was not only published in a peer-reviewed journal [28], but was also registered with the International Prospective Register of Systematic Reviews (PROSPERO) under the registration number CRD42018088599. Furthermore, the reporting of this study adheres to the guidelines outlined by the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) [29].

2.2.2 Eligibility criteria

We employed specific eligibility criteria to determine article inclusion, as outlined below:

Study Design: Articles eligible for inclusion were peer-reviewed studies with study designs encompassing case-control, cohort (prospective or retrospective), cross-sectional, or randomized controlled trials.

Outcome: We focused on studies predicting all-cause in-hospital mortality. Studies that predicted operative, trauma, or post-discharge mortality were excluded.

Setting and Target Population: Our study centered on children over 1 month old admitted to pediatric wards within resource-limited settings, as defined by the World Bank [30]. Studies targeting children in High Dependency Units (HDU) or Intensive Care Units (ICU) were excluded due to the limited availability of such facilities in LMIC. Additionally, studies that included conditions uncommon in children, such as diabetes, cancer, chronic kidney disease, musculoskeletal disorders, etc., were excluded. However, studies focusing on prevalent childhood illnesses, such as malaria, pneumonia, meningitis, anaemia, and diarrhoea/dehydration [3], were included.

Prognostic Research Studies: Studies primarily aimed at developing predictive models or scoring systems were included. Excluded were case series, conference proceedings, editorials, commentaries, expert views, case reports, reviews, and studies primarily generating hypotheses, such as explanatory studies [31].

Predictors in the Model: Studies that reported multivariable models with a minimum of 2 variables/predictors were included.

Full Text and Language: Language restrictions were not applied, and non-English language studies were translated using Google Translate. Studies not available in full text were excluded.

2.2.3 Search strategy of articles

Following the CHARMS (Checklist for critical Appraisal and data extraction for systematic Reviews of prediction Modelling Studies) checklist [32], our study identified seven core items, detailed in Table 2.2.3-2 which provided precise guidance for the formulation of our eligibility criteria, review objectives, and the structuring of our search strategy. In our quest to pinpoint relevant research papers that develop predictive models, we made use of MeSH (Medical Subject Headlines) terms and keywords, thoughtfully presented in We searched through CINAHL (via EbscoHost), Google Scholar, MEDLINE, and Web of Science, with our search scope spanning from the inception of these databases to August 2019. To uncover additional studies that might meet our eligibility criteria, we manually scrutinized the reference lists of the identified articles. To consolidate our search results, we effectively employed the EndNoteX7™ bibliography tool.

Table 2.2.3-1. We searched through CINAHL (via EbscoHost), Google Scholar, MEDLINE, and Web of Science, with our search scope spanning from the inception of these databases to August 2019. To uncover additional studies that might meet our eligibility criteria, we manually scrutinized the reference lists of the identified articles. To consolidate our search results, we effectively employed the EndNoteX7™ bibliography tool.

Table 2.2.3-1: Systematic review framework as recommended by CHARMS checklist.

Item	Criteria
Prognostic or diagnostic model	Prognostic model predicting in-hospital mortality.
Scope	Prognostic models to inform clinicians about the risk of deterioration or death.
Type of prediction models	Prognostic models with and/or without external validation.
Prediction target population	Children aged > 1 month to 15 years admitted in pediatric wards in developing countries
Outcome of interest	All-cause in-hospital mortality.
Prediction period	Any
Intended moment to apply the prediction tool	Prognostic model to be used in primary prevention to assess risk of deterioration and thus guide prevention/treatment.

CHARMS= Checklist for critical Appraisal and data extraction for systematic Reviews of prediction Modelling Studies

Table 2.2.3-2: Search terms for prognostic models

Search ID	Sub-heading	Search Terms
S4	Children	paediatric* OR pediatric* OR (MH "Pediatrics+") OR child*
S3	Hospital based	(MH "Hospitals+") OR hospital*
S2	Low-income countries	(MH "Developing Countries+") OR (MH "Africa+") OR TI ("low income" OR "low and middle income" OR "LMIC" OR "LIC" OR "limited resource*" OR "poor resource*" OR "resource* poor" OR ("developing countries") OR ("developing nations") OR ("third world") OR "resource-constrained" OR ("global south"))
S1	Predictive models	prognos* OR (MH "prognosis") OR (Predict* AND (Outcome* OR Risk* OR Model* OR Mortality OR Index OR Rule* OR decision* OR scor*)) OR "risk score" OR "scor* system" OR "logistic model*" OR "risk prediction" OR "risk calculation" OR "risk assessment" OR "c statistic" OR discrimination OR calibration OR AUC OR "area under the curve" OR "area under the receiver operator characteristic curve"

2.2.4 Data extraction from the included articles

In our analysis of each study within our scope, we adhered to the guidelines outlined in CHARMS, ensuring the extraction of a comprehensive set of data points. This data encompassed particulars like the enrollment of participants, study design, characteristics of the study population, geographic location, sample size, the number and selection of predictors, study duration, methods for handling continuous predictors, management of missing data, modeling techniques (such as logistic regression or survival analysis), validation of model assumptions, internal validation methods (e.g., cross-validation, bootstrapping, or random data splits), presentation formats (such as regression formulas with coefficients, score charts, or nomograms), and model performance metrics.

These performance metrics embraced aspects of discrimination, quantified by measures like the area under the curve (AUC) complete with 95% confidence intervals (CIs), as well as calibration and classification metrics, including specificity, sensitivity, positive predictive value, and negative predictive value. Beyond this, we delved into an extensive literature review to ascertain whether the models under consideration had undergone external validation in other studies. In cases where an article featured multiple prognostic models, each model was independently scrutinized. Importantly, to ensure the utmost accuracy and consistency, the data extracted by our two reviewers underwent a meticulous cross-checking process, with any disparities being resolved through thoughtful discussion with a third reviewer.

It's worth noting that due to substantial heterogeneity among the studies included in our analysis, we opted not to engage in a quantitative synthesis of the identified models.

2.2.5 Assessment of methodological rigor of the identified prognostic models

Following the PROBAST framework (Prediction study Risk of Bias Assessment Tool), a Cochrane tool meticulously designed for assessing the risk of bias (RoB) in predictive models[33, 34] we carried out RoB evaluations for each model across four pivotal domains. These domains encompassed:

- a) **Selection of Study Participants:** This domain probed how participants were recruited for the study.
- b) **Predictors Domain:** Here, we delved into elements like the selection of candidate predictors.
- c) **Statistical Analysis Domain:** This involved the evaluation of factors such as sample size, treatment of continuous predictors, and the management of missing data.
- d) **Outcome Domain:** This focused on the methods used for measuring and reporting outcomes.

Within each of these domains, we employed a series of signaling questions, each offering five potential responses: yes, probably yes, probably no, no, and no information. Any affirmative response (yes or probably yes) signaled a low RoB. The final RoB rating for each model was determined by the culmination of outcomes within these domains.

In alignment with PROBAST recommendations, a prognostic model was stamped with a "low RoB" rating if all four domains were marked as "low" RoB. Conversely, a model was tagged with a "high RoB" when at least one domain received a "high" RoB rating. Models teetered into the "unclear RoB" category if at least one domain garnered an "unclear" rating, while the remaining domains maintained a "low" RoB rating.

2.3 Results

2.3.1 Characteristics of the included studies

Our search strategy initially unearthed a total of 4054 unique articles. However, 3545 of these were subsequently excluded after a thorough review of titles and abstracts, as they veered into non-relevant topics. The full texts of 509 articles underwent a rigorous assessment for eligibility, eventually leading to the inclusion of 15 primary studies that reported a total of 21 developed models, all of which met the eligibility criteria. See Figure 2.3.1-1 for a visual representation of this selection process. The eligible studies analysed data for patients who were below 15 years of age with median mortality being 6.7% (range 1.2% to 43.9%).[35] [36] While majority of the models were developed for general cases in paediatric wards (n=9), some were tailored for specific paediatric groups defined by common diagnoses such as febrile illness (n=1),[37] malaria (n=2), [38, 39] pneumonia (n=4),[35, 40-42] malnutrition (n=2) [43, 44] and other infectious diseases (n=3).

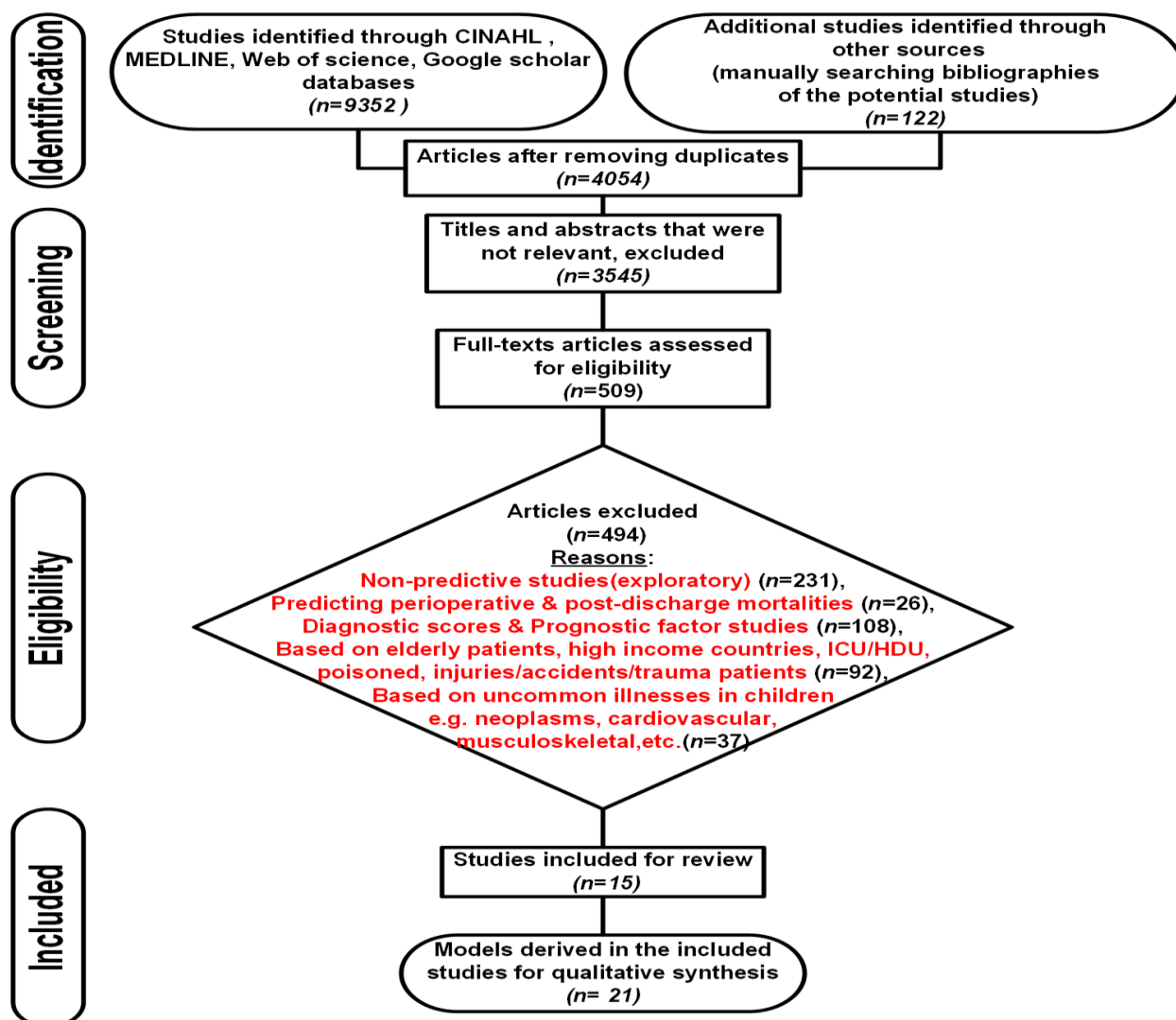


Figure 2.3.1-1: PRISMA flow diagram showing the process used to identify prognostic models predicting in-hospital paediatric mortality included in this review.

The bulk of the studies included in our analysis were published after the year 2000, totaling 20, except for a single study published in 1996 [43]. Notably, the temporal range of data used in these models under review varied, with the latest dataset employed in Rosman et al.'s study spanning from 2016 to 2017 [45]. In contrast, the oldest dataset was utilized by Drimax et al. and encompassed the years 1986 to 1988. Five of the 15 included studies were based on data collected from at least two hospitals. Among these, three studies expanded their research across multiple countries, including countries within sub-Saharan Africa and Asia (Figure 2.3.1-2). In the review of the identified studies, we observed that a significant portion of the essential

information we aimed to abstract was either not reported or was only partially reported. This observation serves as a clear indication of non-compliance with the Transparent Reporting of a Multivariable Prognostic Model for Individual Prognosis or Diagnosis guidelines (TRIPOD) [46, 47].

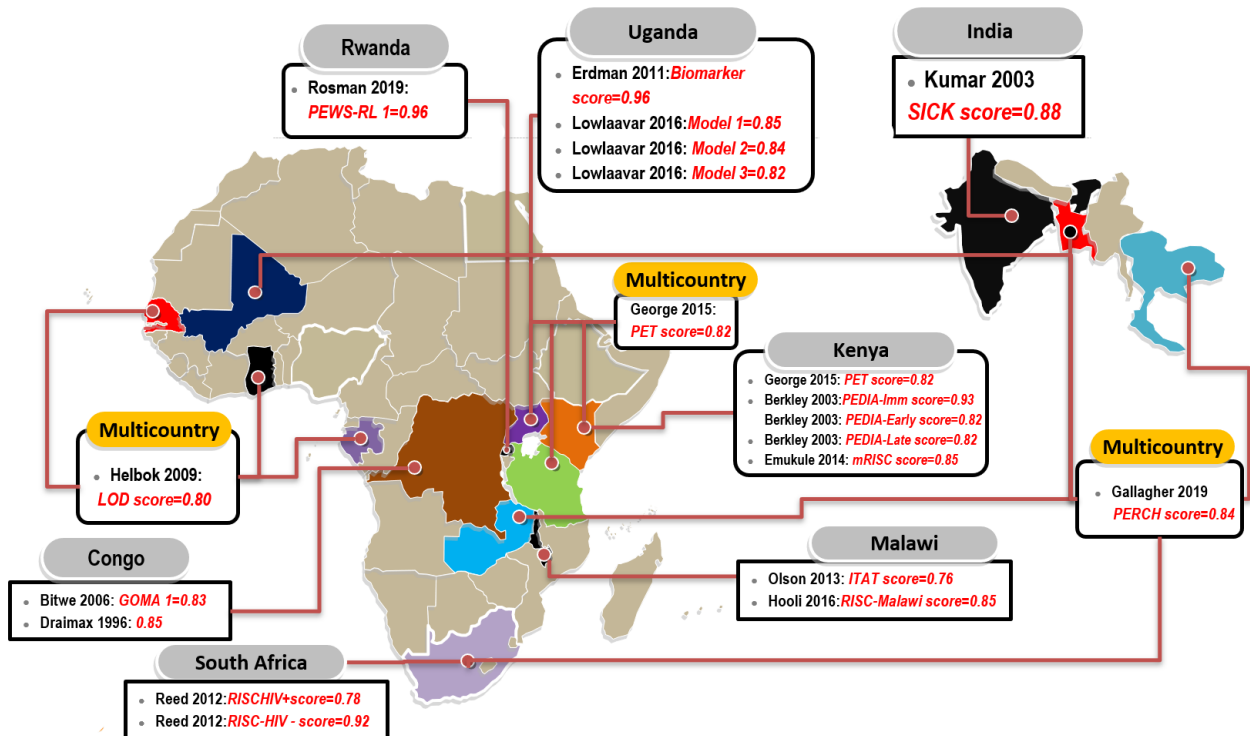


Figure 2.3.1-2: Prognostic models predicting in-hospital paediatric mortality identified by country. Text highlighted in red are the names of the models with their corresponding discrimination measures (area under the curve). Key: PEWS-RL score=Paediatric Early Warning Score for Resource-Limited Settings; SICK score=Signs of Inflammation in Children that Kill; PET score= Paediatric Emergency Triage; mRISC score= Modified Respiratory Index of Severity in Children score; RISC score= Respiratory Index of Severity in Children score; PERCH severity score= Pneumonia Etiology Research for Child Health severity score; LOD score= Lambarene Organ Dysfunction score; CRT= Classification and Regression Trees; ITAT Score= Inpatient Triage Assessment and Treatment score; PEDIA score= Paediatric Early Death Index for Africa score.

2.3.2 Methodological issues of the reviewed models

Candidate predictors

The final reported models featured a total of 61 distinct predictors, with each model typically incorporating a median of 7 predictors. In most cases, the initial selection of independent candidate predictors was predominantly based on univariable analyses. However, it's worth noting that three studies [41] deviated from this approach, opting for predictor selection through literature reviews or considerations of clinical relevance. Across 6 models, a backward stepwise selection method was employed during multivariable analyses to determine the final model predictors. Notably, the common predictors included in the final models encompassed indicators such as altered consciousness, malnutrition indicators, vital signs, and signs of respiratory distress (see Figure 2.3.2-1). Some models included predictors that were either not easily obtainable or required specialized laboratory techniques. Among the 13 models that used continuous predictors, 8 of them categorized these continuous predictors, despite the potential for a continuous scale. Additionally, two out of the 13 models implemented alternative techniques such as fractional polynomials [37] and restricted cubic splines [44] to determine the most suitable functional form for these continuous predictors.

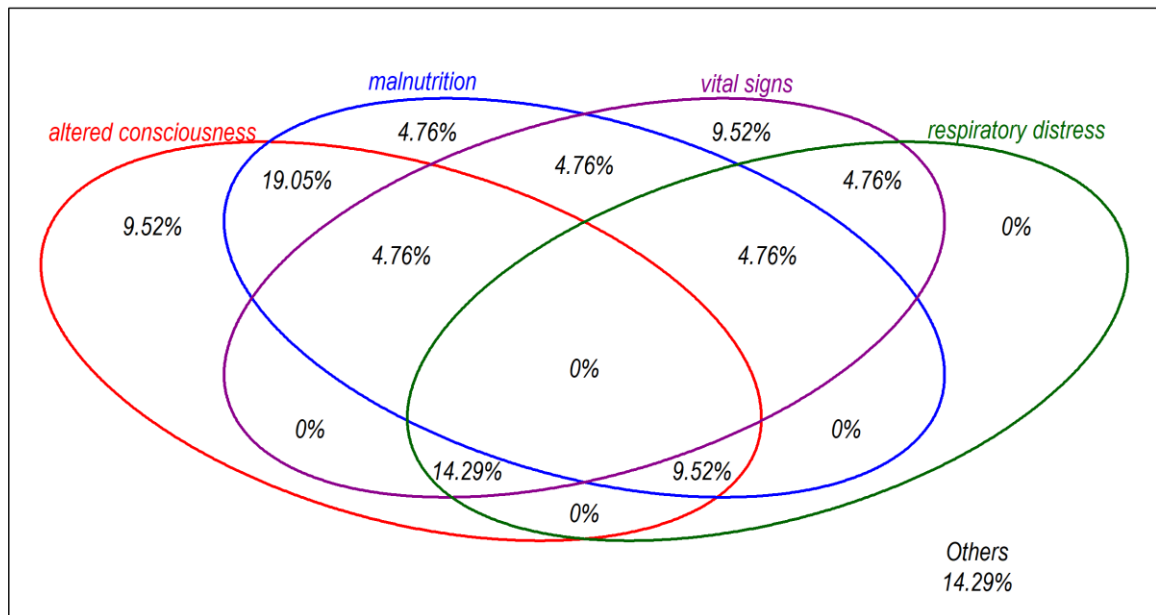


Figure 2.3.2-1: Top four categories of predictors in the models of the reviewed reports: altered consciousness (coma, prostration, not alert, unconscious); malnutrition indicators (kwashiorkor, edema, weight-for-height z-score, weight-for-age z-score, mid-upper arm circumference-MUAC, wasting); vital signs (temperature, respiratory rate, heart rate, oxygen saturation); signs of respiratory distress (indrawing, lung crepitation, difficult breathing, grunting).

Sample size, events per variable (EPV) and missing data

The sample size across the included studies exhibited a wide range, spanning from 16 to 50,249, with a median of 1,307. The median effective events per variable (EPV) stood at 21, with an interquartile range between 8.3 and 32.5. Notably, 7 models had an EPV of less than 10, indicating the potential for overfitting due to insufficient sample sizes. For instance, in the development of the PEDIA-Immediate score by Berkley et al., the dataset reported 60 deaths. According to the general guideline that a study building a predictive model should ideally have a minimum of 10 events (deaths) for each independent candidate predictor,[48] this should have led to the consideration of a model with a maximum of 6 predictors. However, 10 predictors were incorporated instead, resulting in an EPV of 6.

In the case of missing data, there was a lack of uniform reporting. Different approaches were employed across the reviewed studies: 6 models did not report how they handled missing data; 8 employed complete case analysis; 4 utilized multiple imputations through chained equations; and a single study [44], applied single imputation.

Model development

The majority of studies predominantly employed logistic regression for their model development, while one study [37] opted for Cox regression, another study [36] used the Spiegelhalter/Knill-Jones method, and a fourth study [39] utilized a machine learning technique known as classification and regression trees. In the context of verifying model assumptions, most studies failed to provide such information. For example, George *et al.* [37] who employed Cox regression, did not report the verification of the proportional hazard assumption or explore the potential presence of competing risks, as recommended [49]. Other regression assumptions, such as multicollinearity, were also largely unreported. However, it's worth noting that the use of a backward elimination method inherently addresses redundant variables, implying the satisfaction of the multicollinearity assumption if this method was applied [50]. Furthermore, while five studies developed models using data from different countries or centers, none of them clustered their analyses based on the data source in a multilevel model to account for potential heterogeneity. Neglecting this clustering aspect can introduce bias in predictor effects [51].

Model performance evaluation & presentation

Performance measures (both calibration and discrimination) were poorly reported in most of the studies and in most cases (n=20) AUC for discrimination was reported. Performance of the derived models was evaluated in 12 models using either split-sample, resampling methods, or separate datasets. Except for the model derived by George *et al.* [37], all other models did not report both apparent discrimination (without any adjustment for optimism) and optimism-corrected discrimination measures. Despite inadequate reporting of the models' performance, 16 models reported AUCs ≥ 0.80 , an indication of promising models. Apart from the following exceptions; Lambarene Organ Dysfunction (LOD) score [38], Paediatric Early Death Index for Africa (PEDIA) score [36], Signs of Inflammation in Children that Kill (SICK) score [52], Respiratory Index of Severity in Children(RISC) score [35], and Modified Respiratory Index of Severity in Children (mRISC) score [40], other prognostic models in this review have not been externally validated (by independent investigators using diverse populations). Only 2 studies [41] developing 4 models provided a full model formula (both coefficients and intercept/baseline function) in their results as recommended [46, 47]. While most of the models (n=17) were presented as simplified integer scores, only a few were assigned weights according to the regression coefficients.

Risk of bias (RoB)

According to the PROBAST tool, Risk of Bias (RoB) was assessed in four key domains: participants, predictors, outcome, and analyses. Figure 2.3.2-2 provides a summary of the RoB assessment across all models included in this review. Notably, the domain of outcome showed consistently low RoB across all models. However, the domain of statistical analyses raised significant concerns. In 19 out of 21 models, comprehensive details of model development were not reported as expected, making it challenging to conduct a proper risk of bias assessment using the nine signaling questions under the analysis's domain. Consequently, these models

were judged to have an unclear RoB in this domain (see Figure 2.3.2-3). In the overall RoB judgment, 9 out of the 21 models were assessed as having a high risk of bias because at least one out of the four domains in these models received a high RoB rating. The remaining models (12 out of 21) were judged to have an unclear RoB due to the combination of low and unclear RoB ratings in the domains. Notably, no model received a low RoB rating in all four domains.

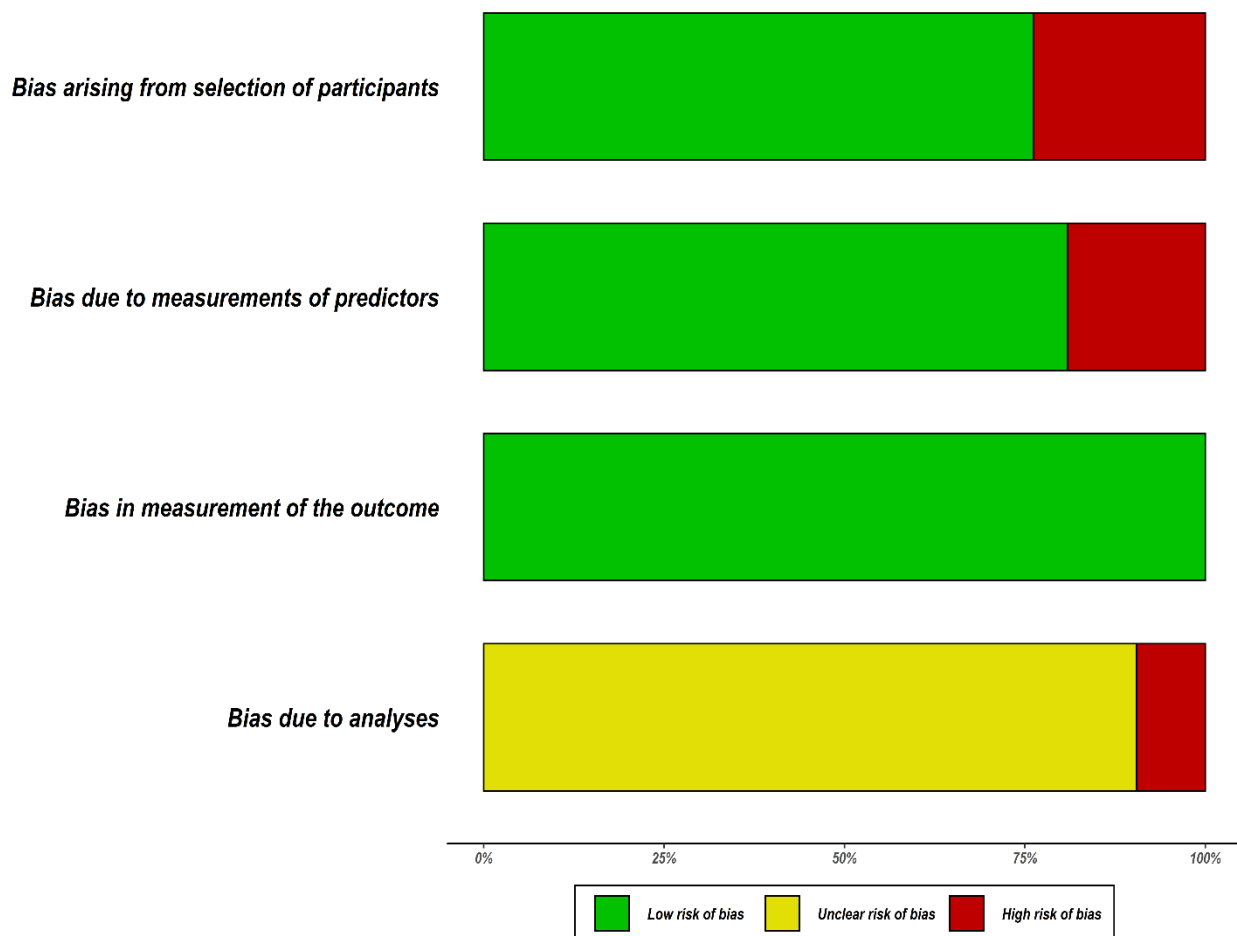


Figure 2.3.2-2: Summary of the risk of bias of the included models using PROBAST (Prediction study Risk of Bias Assessment Tool).

Study	Risk of bias domains				Overall
	D1	D2	D3	D4	
Berkley 2003 (PEDIA -Immediate score)	+	+	+	?	?
Berkley 2003(PEDIA -Early score)	+	+	+	?	?
Berkley 2003(PEDIA -Late score)	+	+	+	?	?
Bitwe 2006 (Goma score)	+	+	+	?	?
Drimax 1996	+	X	+	?	X
Kumar 2003(SICK score)	+	+	+	?	?
Geoge 2015 (PET score)	+	+	+	?	?
Emukule 2014 (mRISC score)	X	X	+	X	X
Reed 2012 (RISC HIV+ score)	+	+	+	?	?
Reed 2012 (RISC HIV- score)	+	+	+	?	?
Hooli 2016(RISC-Score Malawi)	X	+	+	?	X
Gallagher 2019(PERCH Score)	+	+	+	X	X
Helbok 2009(LOD score)	+	+	+	?	?
Erdman 2011(logistic regression)	+	+	+	?	?
Erdman 2011(CRT)	+	+	+	?	?
Lowlaavar 2016 (Model 1)	+	X	+	?	X
Lowlaavar 2016 (Model 2)	+	X	+	?	X
Lowlaavar 2016 (Model 3)	+	+	+	?	?
Mpimbaza 2015	X	+	+	?	X
Olson 2013(ITAT score)	X	+	+	?	X
Rosman 2019(PEWS-RL score)	X	+	+	?	X

Domains:
D1: Bias due to participants selection
D2: Bias due to predictors measurements.
D3: Bias due to determination of outcome.
D4: Bias due to analysis.

Judgement
X High
? Unclear
+ Low

Figure 2.3.2-3: Risk of bias assessment. Low means low risk of bias, High means a high risk of bias, and Unclear bias means it was not possible to assess the risk of bias. Key: PEWS-RL score=Paediatric Early Warning Score for Resource-Limited Settings; SICK score=Signs of Inflammation in Children that Kill; PET score= Paediatric Emergency Triage; mRISC score= Modified Respiratory Index of Severity in Children score; RISC score= Respiratory Index of Severity in Children score; PERCH severity score= Pneumonia Etiology Research for Child Health severity score; LOD score= Lambarene Organ Dysfunction score; CRT= Classification and Regression Trees; ITAT Score= Inpatient Triage Assessment and Treatment score; PEDIA score= Paediatric Early Death Index for Africa score.

2.4 Discussion

2.4.1 Summary of key findings

In this chapter, we conducted a systematic review with the objective of identifying predictive scores for in-hospital mortality among pediatric patients in resource-limited countries. During this comprehensive review, we identified fifteen studies that collectively presented the development of twenty-one distinct prognostic models. Our analysis not only involved an examination of the characteristics of these studies but also an in-depth assessment of the methodological quality of the included models, adhering to contemporary guidelines relevant to predictive models.

This assessment brought to light several significant quality issues. These concerns primarily revolved around reporting deficiencies and various methodological considerations. Key problems included inadequate handling of missing data, a heavy reliance on univariable analysis for predictor selection, inappropriate categorization of continuous predictors, suboptimal adherence to the events per variable (EPV) principle, and less-than-optimal presentation of the proposed models for practical application. As a result, none of the models met the established criteria for good methodological quality, indicating an overall risk of potential high or unclear bias in their predictive capabilities (see Figure 2.3.2-3).

Our analysis of predictive models reveals discrepancies in meeting contemporary methodological standards. Firstly, we observed that 18 out of the 21 models in this review routinely employed univariable analysis for the initial selection of candidate predictors to be integrated into multivariable analyses. This common practice may exclude potentially significant prognostic factors that, while appearing insignificant in univariable analyses, could prove to be substantial when combined with other predictors [46, 47]. It is recommended to make an a priori selection of predictors based on expert opinion, clinical intuition, or relevant

literature for this purpose. However, it's worth noting that only three studies in this review followed this approach [41].

Using small sample sizes during model development can result in poor predictive performance, overfitting, and biased effect estimates. To ensure the reliability of prognostic models, it's generally accepted that there should be a minimum of ten events for each candidate independent predictor [53-55]. Models with insufficient events per variable (EPV) are considered underpowered and are more likely to produce spurious results [48]. Surprisingly, in this review, 7 out of 21 models had inadequate sample sizes ($EPV < 10$), and there was no information provided regarding whether bootstrapping, which helps reduce overfitting, was utilized in these models [56].

Similar to many epidemiological studies, dealing with missing data is a common challenge. The typical approach to address this issue involves multiple imputation or other suitable methods. However, in the model development studies reviewed here, such methods were rarely used. To illustrate, out of the 21 models examined, 8 employed Complete Case Analysis (CCA), 4 used multiple imputation under the assumption that the missing data was Missing at Random (MAR), and 6 did not provide information on how they handled missing data, leading us to assume they used CCA. Following Harrell's recommendations [57], CCA should only be considered when the percentage of missing data is less than 5%. However, the appropriateness of using CCA could not be determined, as most of the studies failed to report the proportion of missing data for each variable. Inappropriately using CCA means that only a small subset of the data is considered, which cannot be considered a random sample from the target population unless the data is Missing Completely At Random (MCAR), a condition that is rarely observed in practice [58]. Consequently, there are concerns regarding potential loss of precision in inferences and biases in the estimated parameters [59] or models employing CCA. Multiple Imputation by Chained Equations (MICE) is the recommended method for handling missing

data, but applying it when data are not missing at random could lead to biased model results [60]. Additionally, there were concerns about how continuous predictors were handled in this review. Among the 13 models that included continuous predictors, 8 of them chose to categorize these variables, even when a continuous scale was feasible. While this approach may seem intuitive, it comes at the cost of predictive accuracy, resulting in poorer model performance due to a loss of statistical power and information.

It is recommended that the nature of continuous data should be retained or managed using appropriate techniques, such as flexible parametrizations like fractional polynomial regression splines, or non-parametric methods like locally estimated scatterplot smoothing (LOESS) functions [61, 62]. In this review, only 2 studies applied appropriate methods for transforming continuous data, using restricted cubic splines and fractional polynomial approaches.

While sixteen models achieved a discrimination metric of over 80%, indicating their promise, it's crucial to exercise caution when interpreting their performance. This caution is warranted because the median mortality rate in the studies included was only 6.7%, resulting in heavily imbalanced data due to the rarity of the outcome of interest. To illustrate this point, consider a study with a mortality rate of 5%; a model predicting no deaths could easily achieve 95% accuracy, which could be misleading [49, 63]. Therefore, it is advisable for authors to provide additional performance measures for their models. These measures may include model specificity, sensitivity, accuracy, positive predictive values, and negative predictive values. These additional metrics will enable a more accurate contextualization of the model's performance, considering the inherent challenges of imbalanced data.

2.4.2 Comparison with other Studies

The methods employed to evaluate the quality measures of the models included in this chapter have previously been used to assess predictive models in various specialties [64-66]. As seen in the findings of this review, earlier reviews [26, 67-69] that focused on the development of

prognostic models also identified numerous shortcomings. These issues encompassed inappropriate statistical analyses, insufficient reporting of critical methodological details necessary for model validation, and a general lack of external validation.

It's essential to recognize that a detailed and transparent report of the methods used in model development is a fundamental principle of research integrity. This transparency enables the research community to assess study findings and gauge the risk of bias [70]. Inadequate reporting of clinical models not only hampers future research in the field of prognostics but also contributes to wasted research efforts [71]. For instance, conducting external validation of prognostic models requires access to a complete model formula. This information is crucial for directly estimating survival probabilities [47]. However, this level of detail was found in only four of the models. Surprisingly, in the case of five models reported to have undergone external validation, the complete model formula, as required, was not provided. This omission raises questions about whether the authors of these external validation studies applied the original model coefficients to the external datasets or if they essentially created new model coefficients, which is tantamount to model redevelopment.

As a result, this review underscores the importance of researchers adhering to the TRIPOD guidelines, which were established to assist authors of prognostic models in creating comprehensive and transparent reports. It's worth noting that the quality of clinical predictive models does not appear to have improved over time, as previous reviews spanning from 1996 [72] to 2019 [73] consistently identified suboptimal methodologies in the development of predictive models, particularly in the realm of analysis. Models derived in a subpar manner can lead to overly optimistic results and potentially misleading performances.

Several factors may contribute to the prevalence of low-quality prognostic models, including the pressure to publish new predictive models, regardless of their clinical value, and the insufficient biostatistical support provided to investigators.

Hence, there is a crucial role for the editorial process in promoting best practices and ensuring compliance with the recommendations outlined in the TRIPOD statement. This compliance could be incorporated into the checklist for submission, thereby fostering the development of high-quality predictive models.

2.4.3 Implications of this review

The development of a prognostic model involves several stages, including development, validation (both internally and externally), impact assessment, and implementation. However, a significant number of the models reviewed are still in the initial development stage. This indicates a prevalent focus among researchers on creating new models, often using similar prognostic factors, while neglecting the crucial steps of validating and enhancing existing prognostic models. This lack of validation and improvement leaves healthcare policymakers uncertain about which models to recommend for use in their specific settings.

To progress prognostic research to the next stage, external validation is essential. In Kenya, for example, there are substantial patient-level datasets available, such as the Clinical Information Network (CIN), which has been accumulated over time from various referral hospitals. These datasets can be leveraged for the external validation of the models identified in this review, allowing for comparative assessments, as recommended by Collins *et al.*, [74]. If necessary, predictive performance can be enhanced by incorporating new prognostic factors.

Furthermore, it's worth noting that a considerable number of the models reviewed rounded the original predictor coefficients to the nearest integer. This practice can impact model performance during external validation due to a loss of predictive accuracy resulting from coefficient rounding [58].

In this chapter, we aim to provide guidance on methodological considerations regarding candidate predictors, as identified in this review. When selecting potential candidate predictors for inclusion in a prediction model, researchers should prioritize those that will be readily

available at the time predictions are made. While some predictors obtained from invasive procedures, such as C-reactive protein, blood gas analyses, or blood or cerebrospinal fluid culture, may offer a higher predictive value for mortality, they might not be practical in resource-limited settings where results may take days to be reported or resources may be insufficient to perform such tests in many hospitals. Therefore, models utilizing such variables may not be practical for clinicians in typical emergency departments in low- and middle-income countries (LMIC).

The practice of screening model candidate predictors based on bivariate associations, using a p-value threshold (typically 0.05), has been strongly discouraged in previous research. Additionally, the common practice of categorizing continuous model predictors is also problematic, as it discards valuable information and often lacks clinical plausibility [49]. Finally, there's a risk of overfitting if a model includes more predictors than the dataset can support. The ratio of events (deaths) to the number of independent candidate predictors has been extensively discussed in methodological literature, and it's recommended that the ratio of events per variable (EPV) should be at least 10 [75, 76].

2.4.4 Strengths and limitations of this review

The article search strategy employed in this chapter successfully identified several potentially eligible studies, making it unlikely that any relevant studies were inadvertently omitted. The quality assessment of the included models was based on contemporary reporting standards and applied consistently to all identified studies.

For example, when there was no mention of internal validation or confirmation of model assumptions in a study, we were unable to determine whether these critical steps in model development were conducted or not. Consequently, models that might have otherwise been considered low risk for bias were classified as either unclear or high risk in each domain. The PROBAST analysis domain included the majority (9 out of 20) of the signalling questions, and

any model within this domain had a higher likelihood of being labelled as high risk as long as there was a single negative (no or probably no) response. This stringent criterion resulted in all models being categorized as either unclear or high risk for bias, thereby preventing us from conducting a meta-analysis.

We acknowledge that our conclusion might change if we were to relax this decision rule to some extent. Nonetheless, we maintain our position that authors should adhere to guidelines for transparent and comprehensive reporting of any proposed prognostic model. Such adherence facilitates the model's external validation and subsequent practical application.

Lastly, it's important to note that we used Google Translate to interpret a study by Bitwe *et al.* [77] from French to English. It is possible that some statistical terminologies were not accurately translated, and certain aspects of the model characteristics might have been lost in the process.

2.5 Conclusion

Well-constructed and thoroughly validated predictive models have the potential to make significant contributions to child survival, particularly in resource-limited countries. In our review, we specifically focused on identifying predictive models for in-hospital mortality among pediatric patients. Unfortunately, we found that none of the models we examined met the criteria for being considered of high quality. This highlights the pressing need to address the identified shortcomings in future prognostic model development by adhering to widely accepted and standardized methodological criteria. It's important to emphasize that most of the models created have not yet undergone the crucial step of external validation. This omission not only obstructs rigorous external validation by other researchers but also undermines the practical applicability of these models. Rather than prioritizing the creation of new prognostic models, researchers should strongly consider conducting comprehensive joint external validation exercises using large datasets collected over extended time periods and from diverse

locations. This approach would facilitate comparisons among models and, when necessary, enable adjustments to ensure their generalizability and effectiveness in real-world scenarios.

Chapter 3

Research Methodology

3.1 Introduction

In this chapter, we present key principles and approaches that will be employed across this thesis report. Section 3.2 outlines the data sources and the study designs utilized for data collection, while section 3.3 elaborates on data management and the handling of missing data. The concepts and methods utilized for external validation are detailed in section 3.4, with model recalibration methods discussed in section 3.5. Additionally, sections 3.6 and 3.7 provide insights into strategies for managing model uncertainty, and addressing competing risks, respectively.

3.2 Data

3.2.1 Data sources

We leveraged on the data collected by the Clinical Information Network (CIN) from 20 Kenyan county referral hospitals. The geographical locations of these hospitals are depicted in Figure 3.2.1-1. The selection of hospitals for inclusion in the CIN dataset was deliberate, focusing on those with an annual admission rate of at least 1000 patients. These hospitals were chosen to provide a representative sample encompassing both high and low malaria endemicity regions, spanning both large rural and urban settings. More details about the selection of these hospitals and their locations have been given elsewhere [78]. The CIN initiative is a collaborative effort involving medical researchers from the Kenya Medical Research Institute (KEMRI)-Wellcome Trust Research Programme, the Kenya Ministry of Health (MoH), the Kenya Paediatric Association (KPA), and the University of Nairobi (UoN). Since 2013, the CIN has been

collecting over 200,000 de-identified patient-level records from pediatric wards, which includes data on more than 6,000 patient fatalities.

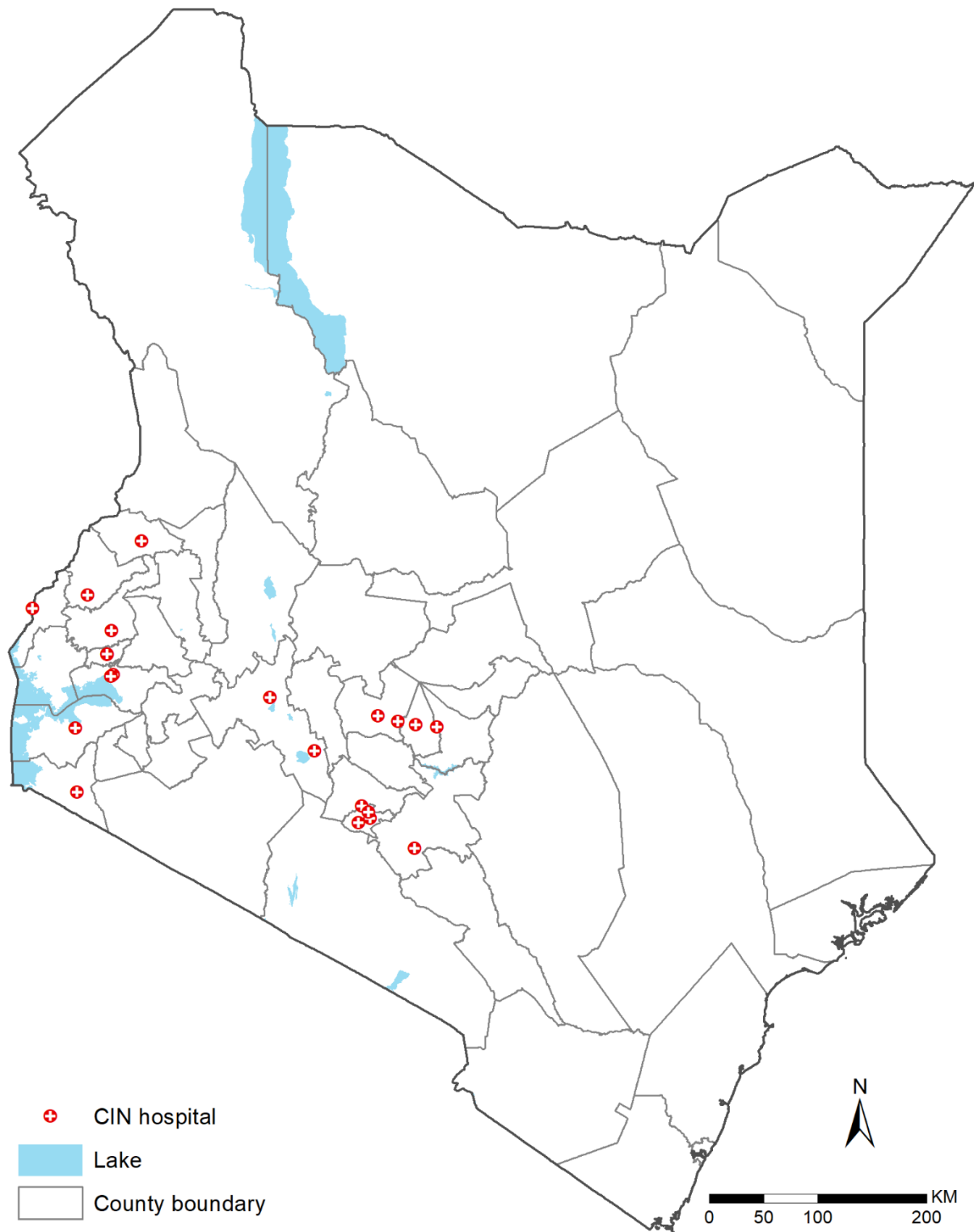


Figure 3.2.1-1: Locations of hospitals included in the validation cohort.

3.2.2 Data management

Patients admitted to hospitals affiliated with the Clinical Information Network (CIN) undergo thorough data collection procedures. This process involves documenting various aspects of their medical history and care. The information routinely collected includes biodata like age and gender, details about the patient's illness history such as the duration of illness, fever history, diarrhea, vomiting, convulsions, and vaccination records. Additionally, examination findings are recorded, covering vital signs, the presence of conditions like thrush or edema, and visible wasting.

Investigation results from tests conducted during admission, such as malaria, hematology, glucose levels, HIV, and lumbar puncture tests, are also documented. Clinician also records primary and secondary admission and discharge diagnoses, as well as the treatments administered during inpatient, including antibiotics, anti-malarial drugs, and anti-tuberculous medications. Any supportive care measures provided, such as oxygen support, blood transfusions, and fluid treatments, are also documented.

Vital signs measurements taken during the initial 48 hours of admission and patient outcomes at the time of discharge are also recorded. To document care provided in hospitals, clinicians in these hospitals use a standardized medical record known as the Pediatric Admission Record (PAR), which is universally adopted for use within the CIN network [79].

After a patient's discharge or death, a trained data clerk transfers this data to a customized data capture tool. This tool is developed using the non-proprietary Research Electronic Data Capture (REDCap) platform [80]. As part of quality assurance procedures, local data quality checks are conducted before data is synchronized with a central database at the end of each day. The checks are done using scripts written in R programming language. These checks assess data completeness and identify any transcription errors, and any inconsistencies or omissions. Any data issue detected is corrected by the data clerk after careful verification with the patient's

medical records. However, it's important to note that the data clerk does not make corrections to documentation errors made by the clinical or nursing teams.

3.2.3 Missing data imputation

Owing to the substantial amount of missing data in the CIN dataset, multiple imputation by chained equations (MICE) was performed to address the problem under the assumption of data missing at random (MAR) mechanism.

According to Rubin's seminal work in 1976, Missing at Random (MAR) implies the existence of a systematic link between the likelihood of encountering missing values and the available observed data, rather than the missing data itself. The likelihood of observation Y_i being missing when considering both Y_i and X_i is equivalent to the probability of Y_i being missing when only X_i is considered as shown in equation (3.3-1)

$$P(Y_{\text{missing}}|Y_i, X_i) = P(Y_{\text{missing}}|X_i) \quad (3.3-1)$$

In this method missing values are imputed using a set of univariate conditional imputation models [81] and generates multiple “complete” datasets with different plausible values of the missing values. As recommended, we included all variables of interest in the imputation model and selected other auxiliary variables in the database giving a total of 53 variables in the imputation model. The auxiliary variables were intended to preserve the relationship among variables [82, 83]. In the imputation model we specified different imputation options conditional on the type of the variable, for instance ordered logistic regression option was applicable to ordinal categorical variables, multinomial logistic regression for nominal multi-level was applicable to categorical variables with more than 2 levels, linear regression for continuous variables, and binary logistic regressions for dichotomous categorical variables. Based on the principle that the number of imputations must at least be equal to the proportion

of incomplete data [82], we generated 70 multiply imputed datasets since 68% the CIN records were incomplete. Graphical comparisons of the kernel density plots of the imputed versus observed values suggested the imputed values were plausible since the distributions of the values from the two datasets (imputed and original) appeared identical as shown in Figure 3.2.3-1. The multiply imputed datasets were used in the subsequent analyses to answer study objectives except for the literature review.

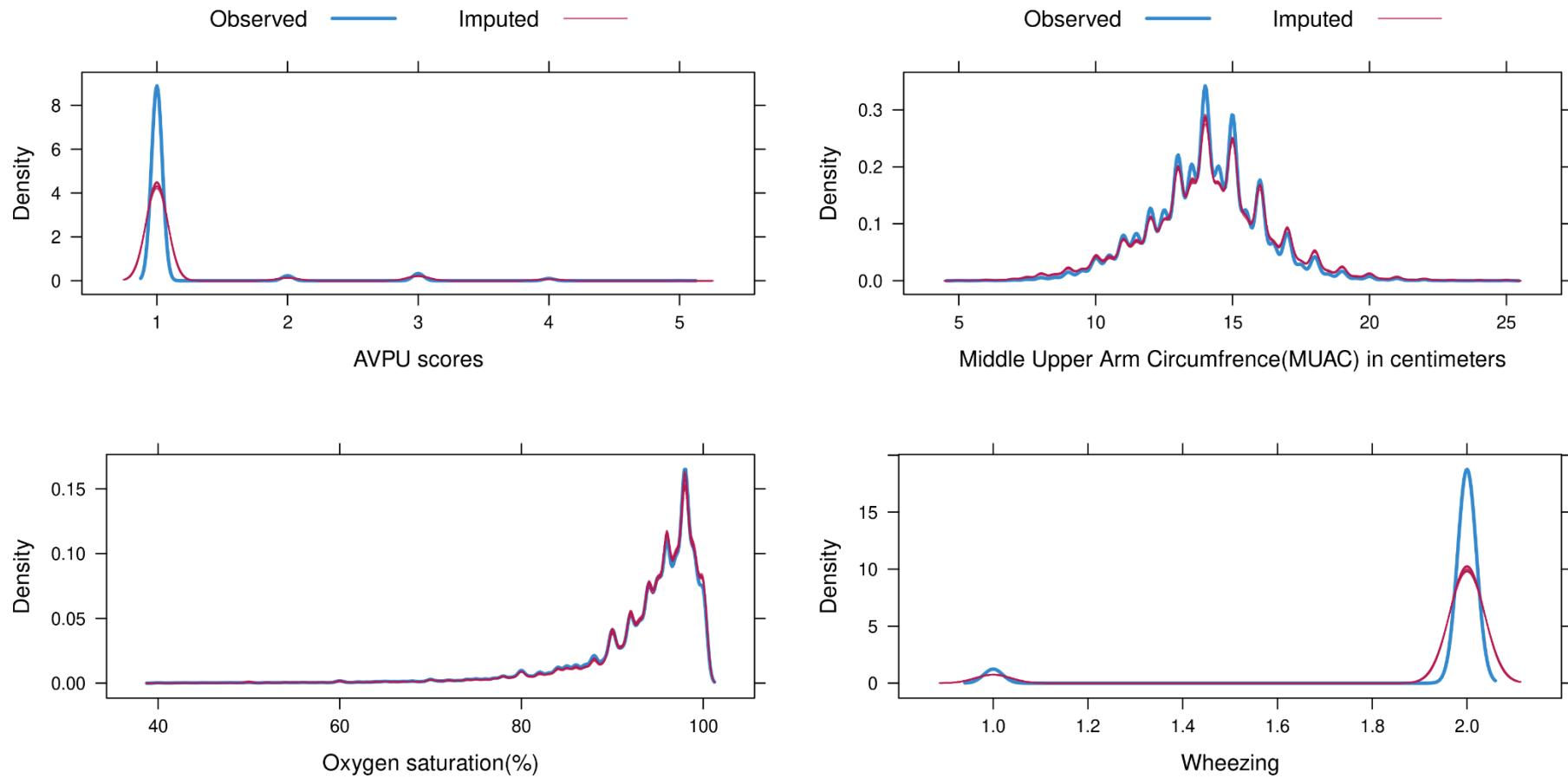


Figure 3.2.3-1: Kernel density plots of the observed (non-imputed) and imputed values of various variables. Visual inspection of the distributions of the observed and imputed values appears identical suggesting the imputation model generated plausible values to replace missing ones.

3.3 Prognostic model's external validations

The literature review conducted in Chapter 2 highlights significant shortcomings in many prognostic models, undermining their practicality and applicability. Common methodological weaknesses include limited sample sizes, resulting in low signal-to-noise ratios and insufficient events-per-variable (EPV), with some models falling below the recommended EPV threshold of 20, potentially introducing bias. Additionally, issues arise from inadequate handling of incomplete data, inappropriate statistical analyses, and overly optimistic interpretations of model outcomes. Excessive dependence on fully automated statistical techniques, like stepwise model selection algorithms (backward or forward), which do not necessitate expert input, can lead to over-optimistic or irrelevant models for real-world use. To determine methodological utility of the identified prognostic models, in Objective II (Chapter 4), we will perform external validation of prognostic models for in-hospital pediatric mortality in LMICs. This validation will be based on routine hospital data collected by the Clinical Information Network (CIN) in Kenya. The following section outlines the methodology to address the research questions in Objective II.

3.3.1 Determining model's predicted risks in the validation set

The model regression coefficients were used to determine predictions of the risk of mortality in the validation dataset. For each patient in the validation cohort, the presence of the model predictor was assigned a value of 1, while its absence was assigned a value of 0. This was then multiplied by the corresponding regression coefficients and added together with the model intercept to get linear predictor. Patient's predicted risk of in-hospital mortality was then computed on the resultant linear predictor using the logistic function provided in equation (3.3.1-1)

$$P(Y = 1|X) = \frac{\exp(\beta_0 + \beta_1 X_1 + \dots + \beta_k X_k)}{1 + \exp(\beta_0 + \beta_1 X_1 + \dots + \beta_k X_k)} \quad (3.3.1-1)$$

where $P(Y = 1|X)$ represents the probability that the dependent variable Y is equal to 1 given the values of the independent variables X . The β_0 is the model intercept and β_i is the regression coefficient for a given predictor X .

3.3.2 Determining model's performance in the validation set

Model performance was determined based on the two metrics namely discriminatory index and model calibration. To determine the discriminatory ability of the model we used the Area Under the Curve (AUC) also known as *c-statistic* which is a measure of the ability of a model/score to distinguish between 2 classes [84, 85]. We classified the model's discriminatory ability using the following cutoffs; $(AUC) \geq 0.90$ was classified as "excellent discrimination", AUC ranging from 0.80 to 0.89 was classified "good discrimination", AUC ranging from 0.70 to 0.79 as "fair discrimination", and "poor discrimination" was for the model whose AUC was < 0.70 [86, 87].

Model calibration was assessed by both plotting the predicted probability of in-hospital death against the observed proportion, and by computing calibration metrics namely *calibration slope* and *calibration intercept* [88]. The *calibration slope*, which has a reference value of 1, examines the dispersion of the predicted risks such that a slope value of < 1 suggests that estimated risks are too extreme while a slope value of > 1 indicates that the estimated risks are too low. On the other hand the *calibration intercept* is a measure of calibration-in-the-large and it has a reference value of 0 such that a *calibration intercept* of < 0 indicate overestimation while that of > 0 indicate underestimation of risk [88]. The confidence intervals for both *c-statistic*, *calibration slope* and *intercept* were calculated through bootstrap resampling using *CalibrationCurves* package in R[89].

3.4 Model recalibration

Clinical prediction models frequently encounter challenges during external validation, resulting in their rejection due to insufficient predictive accuracy. This issue can be partly ascribed to the dynamic nature of clinical settings, where a variety of factors, including changes in clinical practices, influence the landscape despite efforts to standardize these practices through clinical guidelines. Such interventions can modify the prevalence and clinical presentations of common childhood illnesses, diminishing the effectiveness of clinical prediction models developed before these shifts when validated in these evolving settings. Additionally, factors like case-mix variations, disparities in the timing of model development and validation, and dataset drift contribute to the degradation of model performance when applied to new samples. Thus, model recalibration is a vital step to enhance model performance and adapt them to the specific local context, as elaborated in Objective III (Chapter 5). The strategies for model recalibration are elaborated below.

3.4.1 Model recalibration strategies

To improve performance of the existing prognostic models, we used model's linear predictor as shown in equation (3.4.1-1) whereby the α denotes the model intercept and β_1 to β_n denotes the vector of model coefficients (also called slope) for each of the prognostic factor X_1 to X_n (also called covariate).

$$\log\left(\frac{P(\text{in-hospital mortality})}{1 - P(\text{in-hospital mortality})}\right) = \text{linear predictor} = \alpha + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n, \quad (3.4.1-1)$$

The right-hand side of the equation ((3.4.1-1) constitutes the linear predictor ($LP_{original}$) of the original model which is a weighted sum of the prognostic factor X_1 to X_n in the model, weights being β_1 to β_n which are the regression coefficients, and the α denotes the model intercept. This computation is done for each patient meeting the eligibility criteria of the models in the updating dataset. The resultant linear predictor is used by the recalibration strategies to adjust

the model accordingly [90, 91]. In this work, we used recalibration-in-the-large/intercept-only, and logistic calibration method to update the models as described below.

3.4.2 Adjusting model intercept only/recalibration-in-the-large

This approach exclusively adjusts the original model's intercept. The updated intercept is set to match the average of predicted in-hospital probabilities within the dataset used for the update [92]. This was accomplished by fitting a univariate logistic regression model, with the outcome being in-hospital mortality. In this regression, the linear predictor was treated as an offset, effectively fixing a constant coefficient of the covariate at unity for each observation within the updating dataset. Consequently, we derived an intercept from this model, which was then incorporated into the linear predictor of the original model as a correction factor. Importantly, the regression coefficients (β_1 to β_n) of the original model remained unaltered, as illustrated in equation (3.4.2-1).

$$\log\left(\frac{P(\text{in-hospital mortality})}{1 - P(\text{in-hospital mortality})}\right) = (\alpha + \alpha_{\text{correction factor}}) + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n, \quad (3.4.2-1)$$

3.4.3 Logistic calibration

This approach was employed to simultaneously update both the model intercept and the model slope for each of the models undergoing an update. To achieve this, we conducted univariate logistic regressions on each of the updating datasets, with in-hospital mortality as the dependent variable and the linear predictor as a covariate. This modeling process produced two key outputs: the calibration intercept and the slope correction factor. These two quantities were subsequently utilized to refine the original models, as exemplified in the equation 3.4.3-1. This approach effectively updated the original model's regression coefficients proportionally using slope's correction factor.

$$\log\left(\frac{P(\text{in-hospital mortality})}{1 - P(\text{in-hospital mortality})}\right) = (\alpha_{\text{correction factor}}) + (LP_{\text{original}} \times \beta_{\text{correction factor}}), \quad (3.4.3-1)$$

3.5 Accounting for model uncertainty

In the development of prognostic models, researchers often employ selection criteria to choose the "best" model from a pool of competing models, and all subsequent statistical inferences and conclusions are made with the implicit assumption that the distribution function of the selected model accurately represents the actual data-generating model (DGM). This practice is standard in statistical literature. However, it has its limitations, as it neglects model uncertainty in the process of selecting the most appropriate probability distribution function, potentially leading to misleading statistical inferences and either overestimation or underestimation of the risk associated with the outcome of interest. Objective IV (Chapter 6) aims to enhance the predictive ability of the model by addressing model uncertainty. To achieve this objective, the following methodology will be explored.

3.5.1 Stacking of predictive distributions

We started by first fitting models to the data using Bayesian inference. After which we calculated model's posterior predictive distributions which is the distribution of predictions that the model would make for new data. To combine the predictive densities, we used a loss function called Kullback–Leibler divergence to determine model weights such that the model with lowest loss is given the most weight [93].

3.6 Competing risk framework and simulations

All prognostic models discussed in Objectives 1 to 4 focus on pediatric in-hospital mortality, typically occurring shortly after admission, with an average time frame of around 48 hours.

During the development of these prognostic models, the use of the Sub-distribution Hazard (SH) model in scenarios characterized by competing events, short follow-up period or substantial censoring has been discouraged due to its impact on the proportionality assumption of the SH model. Furthermore, it has been argued that, in situations with a brief follow-up period, the competing risks framework has limited influence compared to other alternative approaches. Therefore, Objective V (Chapter 7) aims to investigate how the accuracy of these prognostic models is affected in setups characterized by a short follow-up period or heavy censoring. The following section outlines the methodologies employed in this investigation.

3.6.1 Overview of competing risks framework

When a patient is admitted to a hospital, there are four possible outcomes: death, discharge, referral to another health facility, discharge against medical advice or absconding from the facility. The occurrence of any of these outcomes precludes another event at that point in time. Let $(X_t)_{t \geq 0}$ representing the state a patient is in at any given time, $X_t \in \{1, 2, 3, 4\}$. $X_t = 1$ represent discharge from hospital upon achieving clinical stability, $X_t = 2$ indicates referral to other hospitals for advanced care, $X_t = 3$ indicates the state of being discharged against medical advice or a patient absconds (escapes medical premise without permission of the healthcare worker), and $X_t = 4$ represents in-hospital mortality which is the event of interest. The event-specific (also called cause-specific) hazard functions of these events are denoted by $\alpha_{0j}(t)$, where $0j$ denotes a transition from the point of admission to any of the four competing events at a given time. The competing risk framework is represented schematically in Figure 3.6.1-1 whereby the occurrence of states 1-3 precludes state 4 (in-hospital mortality).

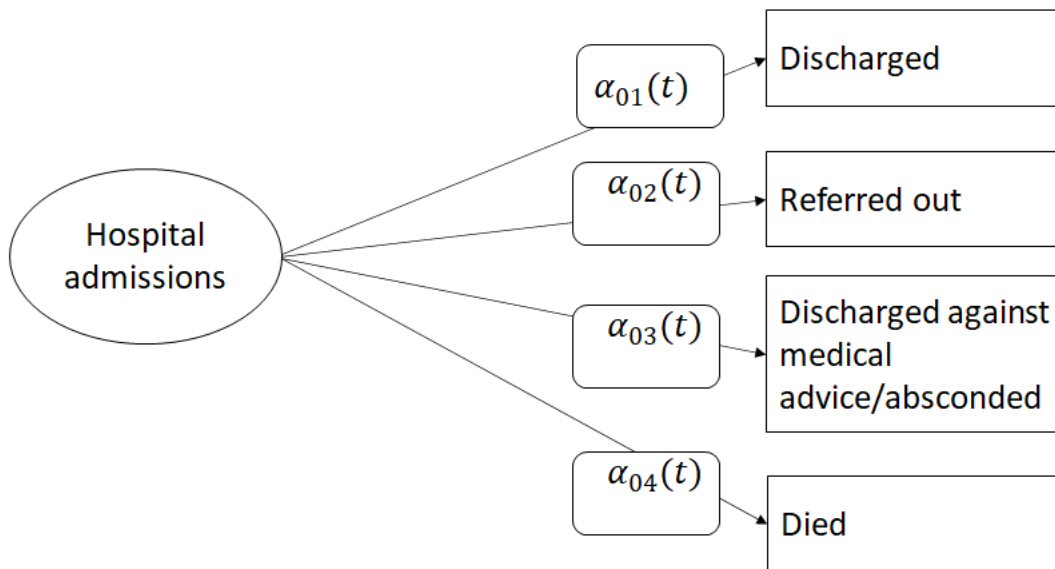


Figure 3.6.1-1: Schematic view of the competing event framework for hospital admissions. Every admitted patient experience any of the four mutually exclusive events (discharged, referred out, discharged against the advice or absconded, and died). Cause-specific hazard functions are denoted by $\alpha_{0j}(t)$, $j = 1, 2, 3, 4$ and each arrow denotes a transition to a state denoted by a rectangle.

3.6.2 Patient survival/follow-up period

The patient survival/follow-up period was determined by length of hospital stay (LOS), which is the difference between the date of the observed event (e.g., date of discharge) and the date of admission. The CIN data had median LOS of 5 days with an interquartile range of 2 to 6 days. Further exploration of the data suggested that a smaller number of patients with LOS >15 days had an outcome of interest (hospital mortality). For this reason, the datasets used in the simulations included patients with LOS ≤ 15 . All simulations and statistical analyses were conducted using the R statistical programming language. The *cmprsk* R package (version 2.2-9) [94] was used to fit the SH models.

Chapter 4

External Validation of Prognostic Models

4.1 Introduction

Childhood mortality remains high in lower and middle-income countries (LMICs) despite a significant reduction since 1990 but it's uncertain how much in-hospital mortality has changed over this duration [95-97]. Mortality rates among hospitalized children in sub-Saharan Africa remains high and most deaths occur within the first few hours of admission [78]. Emergency Triage Assessment and Treatment (ETAT) guidelines, produced by the World Health Organization (WHO), provide guidance on immediate care for children admitted to hospitals. ETAT guidelines provide guidance on triage and uses syndrome-based approach to management of common childhood conditions but they are still used inconsistently and implemented sub-optimally in Africa despite having been in existence since 2005 [98]. Regardless of underlying condition, children at risk of death during hospitalization often present with similar danger signs and prompt triage and immediate supportive management are thought as most important in reducing mortality and morbidity in admitted children[99]. Identification of children at risk of in-hospital mortality is the first step in directing supportive treatments that have the potential to reduce deaths. Therefore, clinical prediction models that identify the sickest children immediately on arrival to hospital for immediate supportive care and targeted close monitoring may be useful [25].

In Chapter 2, It was shown that many prognostic models do not meet the methodological standards hence reducing their utility and generalizability. A common methodological

weakness identified include small sample size which makes a resultant model to have a low signal-to-noise ratio, limited number of events-per-variable (EPV), with some having EPV of less than 20 which is thought to lead to biased estimates [47, 100-102]. Other weaknesses include poor handling of incomplete data, inappropriate statistical analyses, and optimistic interpretations of the model output. Furthermore, overreliance on fully automated statistical techniques, such as the stepwise model selection algorithm (backward or forward), which do not require expert or consensus input, can lead to overoptimistic or biased models that are not always relevant in routine practice [12, 101, 103, 104].

In this Chapter, we focus on the external validation of prognostic models identified in Chapter 2. This is an important step in the development of any prognostic models since it aims to assess predictive model's transferability and or generalizability to other patient populations before it can be recommended for use in the clinical practice. It is noteworthy that in Chapter 2 it was shown that none of the identified model had undergone an independent external validation hence there is uncertainty in their reliability and generalizability [12, 105, 106].

The remainder of this chapter is structured as follows; sections 4.2 present model validation methods, it also delves into the details of the models set to be externally validated as we as sensitivity analysis. The results of this model validations are shown in section 4.3, discussion is provided in section 4.4, implication and conclusion is provided in the section 4.5.

4.2 Methods

4.2.1 Prognostic models included for external validation.

Models included in this external validation chapter were obtained from the systematic review on prognostic models done in chapter 2 which identified 11 models predicting in-hospital mortality for children admitted in LMICs hospitals. However, we excluded 9 models [35-38, 44, 107] since they did not publish the relative weights of the risk factors and model intercept

for the logistic regression models, or the baseline hazard function for the survival models as required for the external validation. Before we made the decision to exclude these models, we contacted the corresponding authors of those studies by email asking for the full model formula without success. The following are the models included in the current external validation study.

1. RISC-Malawi prognostic model

RISC-Malawi is a Respiratory Index of Severity in Children (RISC) that was developed using prospectively collected clinical data from a cohort of 14,665 hospitalized children aged 2-59 months with pneumonia in Malawi between 2011–2014. Total deaths in the model development cohort were 465 deaths and the case fatality rate was 3.2% across the seven hospitals under study[41]. The authors utilized logistic regression to develop a prognostic model whose intercept and odds ratios for the seven prognostic factors are provided in Table 1. The author reported an area under the receiver operator characteristic curve (ROC) of 0.79 (95% CI: 0.76–0.82), demonstrating a fair ability to discriminate between children’s risk of mortality.

2. Lowlaavar et al. 2016 prognostic models

Lowlaavar *et al.* 2016 [108] developed three models utilizing a two-site prospective observational study in Uganda which enrolled children between 6 months and 5 years admitted with a proven or suspected infection. In their study, 1307 children were enrolled consecutively and 65 (5%) of participants died during their in-hospital stay. The study was conducted between March 2012 and December 2013. The primary model included weight for age z-score, Blantyre coma scale and HIV status. Based on the derivation dataset the AUC was 0.85 (95% CI 0.80–0.89). The second model included MUAC (mid-upper arm circumference), Blantyre coma scale, and HIV status. The area under the ROC curve of this model was 0.84 (95% CI 0.79–0.89). Model 3 included 2 variables MUAC and Blantyre coma scale with an AUC of 0.82(0.72-0.91). The equations of these 3 models are provided in Table 4.2.1-1.

Table 4.2.1-1: Models to be externally validated.

Study	Models	Inclusion criteria	Predictors	Model equation with intercept and odds ratios	Model derivation AUC (95% CI)	CIN sample size eligible for validation (% with mortality outcome)
<i>Hooli et al. 2016</i>	RISC-Malawi Model	age \geq 2months <=59months, Pneumonia by danger signs (Cough or difficult breathing and at least one danger sign (Central cyanosis, grunting, chest wall indrawing, stridor, inability to drink, AVPU<A, or convulsion))	Moderate hypoxemia, severe hypoxemia, moderately malnourished, severely malnourished, child- sex(female), wheezing	= -4.67 + (<i>moderate hypoxemia</i> \times 0.43) + (<i>severe hypoxemia</i> \times 1.62) + (<i>moderately malnourished</i> \times 0.55) + (<i>severely malnourished</i> \times 1.53) + (<i>female sex</i> \times 0.22) + (<i>wheeze</i> \times -0.35) + (<i>unconsciousness</i> \times 1.74)	0.79 (95% CI: 0.76– 0.82)	N= 50,669, Mortality=4406(8.7%)

<i>Lowlaavar et al. 2016</i>	Primary model	age \geq 6months <=60months with proven or suspected infection	Abnormal Blantyre Coma Score, Positive HIV, Weight for age z-score	= -4.280 + (<i>abnormal Blantyre Coma Scale</i> $\times 2.51$) + (<i>Positive HIV</i> $\times 1.32$) + <i>weightforage_zscore</i> $\times -0.2$	0.85 (95% CI 0.80– 0.89)	N=86,784 Mortality=4,045(4.7%)
	Model 2.	age \geq 6months <=60months with proven or suspected infection	Abnormal Blantyre Coma Score, Positive HIV, Middle Upper Arm Circumference (MUAC)	= -0.523 + (<i>abnormal Blantyre Coma Scale</i> $\times 2.54$) + (<i>Positive HIV</i> $\times 2.27$) + (<i>MUAC</i> $\times -0.03$)	0.84 (95% CI 0.79– 0.89)	N=86,784 Mortality=4,045(4.7%)
	Model 3.	age \geq 6months <=60months with proven or suspected infection	Abnormal Blantyre Coma Score, MUAC	= 0.303 + (<i>abnormal Blantyre Coma Scale</i> $\times 2.47$) + (<i>MUAC</i> $\times -0.03$)	0.82(0.72- 0.91)	N=86,784 Mortality=4,045(4.7%)

4.2.2 Model validation dataset

Patients hospitalized in paediatric wards aged ≥ 2 months but ≤ 15 years were eligible for inclusion from September 2013 to December 2021 and this included data collected from 20 CIN hospitals. Surgical cases, burns patients, healthy children accompanying sick babies, children admitted with poisoning such as organophosphate ingestion or any other form of poisoning, traumatic and road traffic cases were all excluded from validation cohort. We also excluded patients admitted during healthcare workers(nurses and doctors) strike [109]. These exclusions were done to make the validation dataset as much similar as possible to the derivation cohort of the models whose performance is assessed in this study. To obtain model-specific cohort for the external validation, the following eligibility criteria were applied.

1. Eligibility criteria for RISC-Malawi model's external validation cohort

As defined in the study that derived RISC-Malawi model, the external validation cohort included patients aged 2 to 59 months with admission diagnoses of pneumonia that was defined as follows; history of cough or difficult breathing and at least one of the danger signs which included central cyanosis, grunting, chest-wall indrawing, stridor, inability to drink/breastfeed, and or painful responsive (P) or unresponsiveness (U) based on the disability scale of AVPU (Alert, Verbal, Painful responsive, unresponsive) see Figure 4.2.2-1.

RISC-Malawi model predictors were defined in the validation cohort as follows; moderate malnourished was defined as MUAC between 11.5cm and 13.5cm. Severe malnourished was defined as MUAC<11.5cm. Unconsciousness was assessed using the disability scale of AVPU (Alert, Verbal, Pain, Unresponsive). Thus, a patient was assumed to be unconscious if he/she either responding to pain only or was unresponsive altogether (P or U). Moderate hypoxemia was defined as oxygen saturation ranging from $\geq 90\%$ to $\leq 92\%$, and severe hypoxemia was defined as oxygen saturation $< 90\%$.

To understand performance of RISC-Malawi model in scenario where the definition of a predictors varied from the original study, we performed sensitivity analyses where pneumonia diagnosis was defined based on the clinical diagnosis as opposed to using danger signs.

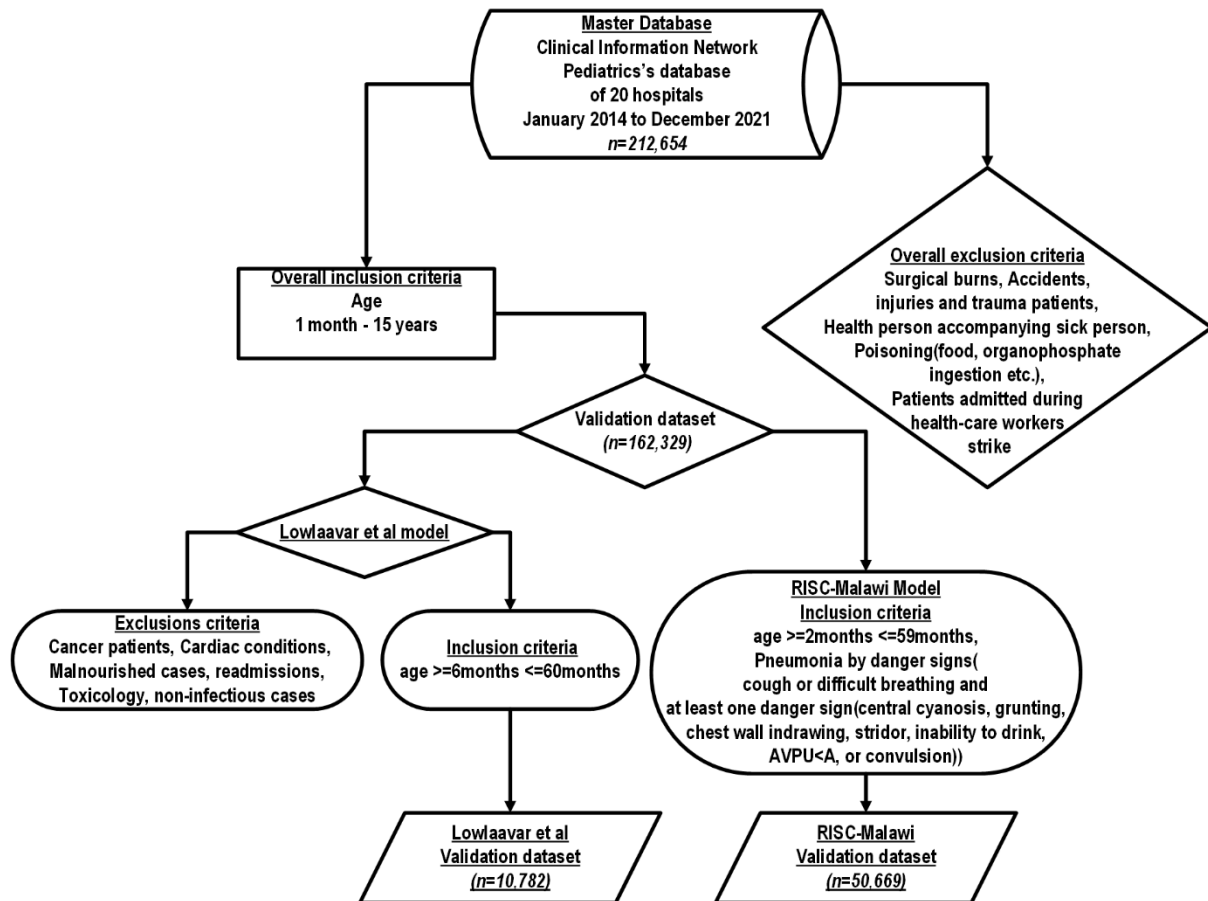


Figure 4.2.2-1: Patients meeting the eligibility criteria of inclusion for external validation of various models.

2. Eligibility criteria for Lowlaavar et al. models' external validation cohort

To match the clinical characteristics of external validation cohort to that of the model derivation, we included children aged 6 to ≤ 60 months and excluded patients with the following characteristics: malnourished cases (defined as clinical diagnoses of malnutrition), readmission cases, those with cancer diagnosis, those with heart condition, and patients with any parasitological confirmed or clinically suspected non-infectious illness see Figure 4.2.2-1.

Model predictors were defined as follows; weight-for-age z-score was computed based on the reference materials in the WHO website (for patients <24 months) [110, 111] and national center for health statistics (for patients >24 months) [112], and abnormal Blantyre coma score(BCS).

Since BCS data was only collected in 6 out of 20 CIN hospitals, we limited the validation data to include patients from the 6 hospitals whose locations are shown in Figure 4.2.2-2. The collection of BCS data was introduced in September 2019 in the 6 hospitals which are participating in the WHO-led study conducting evaluation of pilot subnational introduction of the RTS,S/AS01 malaria vaccine in western Kenya- a region with high malaria transmission throughout the year [113]. Therefore, as defined in the model derivation study, a patient with a BCS of less than 5 was considered abnormal.

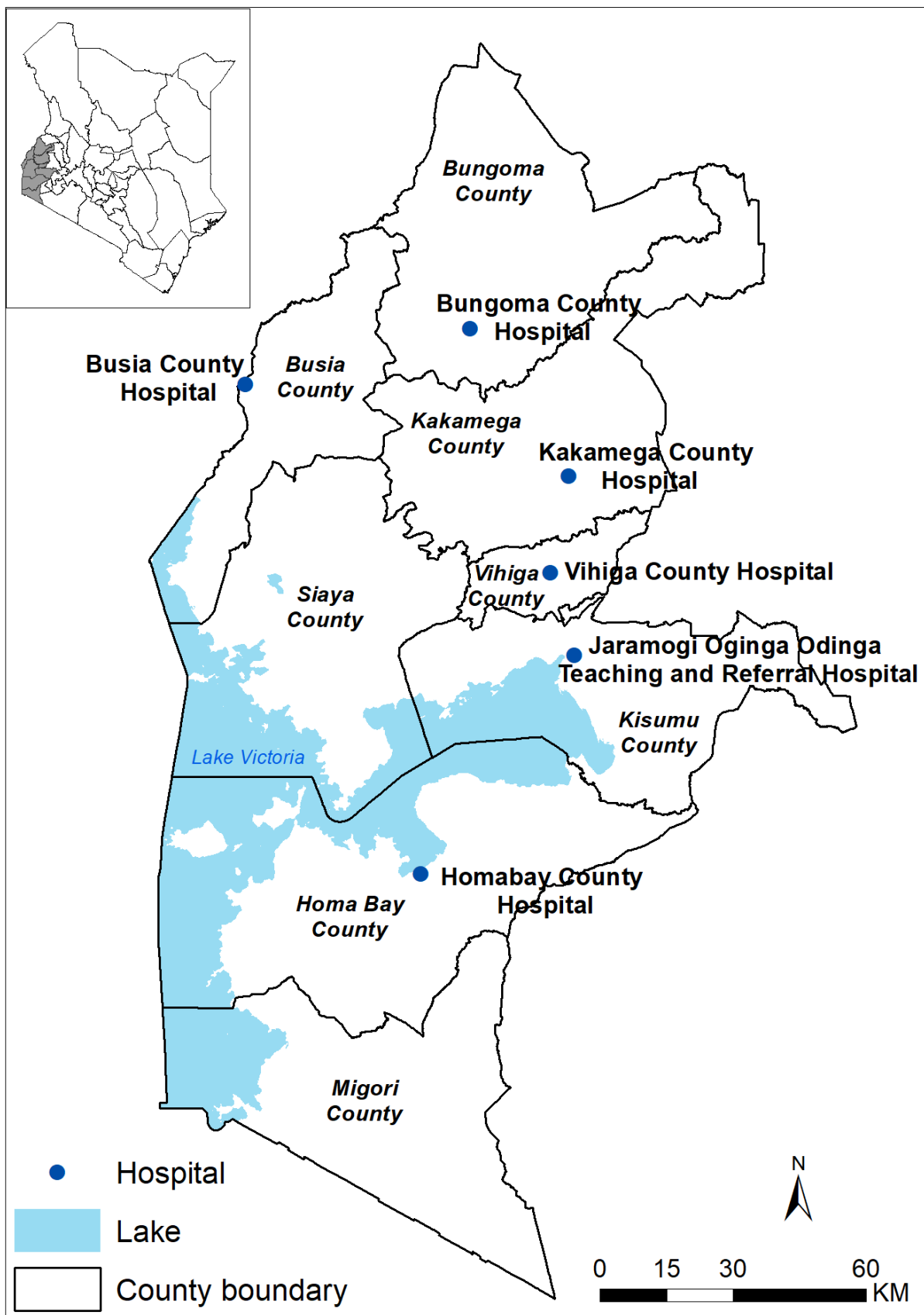


Figure 4.2.2-2: Locations of the 6 hospitals whose patients were included in the validation cohort of Lowlaavar et al. models.

4.2.3 Assessing performance of the prognostic model in the external validation.

Regression coefficients from the model were utilized to predict mortality risk in the validation dataset. The process involved assigning a value of 1 to the presence of the model predictor for each patient in the validation cohort and 0 for its absence. These values were then multiplied by the respective regression coefficients and summed with the model intercept to generate a linear predictor. Subsequently, the patient's predicted risk of in-hospital mortality was calculated based on this linear predictor using the logistic function described in equation (3.3.1-1).

4.2.4 Missing data in the model validation set.

The two models selected for external validation in this Chapter used variables that had varying levels of documentation in the validation cohort, for instance the data in Mid-Upper Arm Circumference (MUAC) which was used to determine malnutrition status were missing in 49.8% of the of the eligible population for RISC-Malawi model. Documentation of this variable was equally poor in the derivation dataset Hooli *et al.* whereby the data were missing in 45.8% of the eligible population. See Table 4.2.4-1 for the data missingness of predictors in RISC-Malawi model and Table 4.2.4-2 for Lowlavaal *et al.* models.

Table 4.2.4-1: Predictors used in RISC-Malawi model and their level of missingness both in derivation and validation datasets.

Predictor	Variable in model derivation dataset	Variable equivalent in external validation dataset	N in the Derivation datasets	N in the Validation datasets
Oxygen saturation	Normal	Oxygen saturation 93%-100%	10,586(64.3%)	16,897(33.3%)
	Moderate hypoxemia (90%-92%)	Oxygen saturation 90%-92%	1,382(8.4%)	3,875(7.6%)
	Severe hypoxemia	Oxygen saturation <90%	2094(12.7%)	8949(17.7%)

	Missing oxygen saturation		2413(14.7%)	20,947(41.3%)
Malnutrition based on Middle Upper Arm Circumference (MUAC)	Normal	MUAC>13.5cm	4557(27.7%)	15,234(30.1%)
	Moderately malnourished	MUAC (11.5cm - 13.5cm)	3382(20.5%)	8,699(17.2%)
	Severely malnourished	MUAC < 11.5cm	991(6.0%)	3,042(6.0%)
	Missing MUAC data		7545(45.8%)	25,232(49.8%)
Wheeze	Wheezing =Yes	Wheezing =Yes	4117(25.0%)	6666(13.2%)
	Wheezing =No	Wheezing =No	8767(53.2%)	42,701(84.3%)
	Missing Wheezing data	Missing data	3591(21.8%)	1,302(2.6%)
Unconsciousness	Unconscious=Yes	Painful responsive or unresponsive in the disability scale of AVPU (Alert, Verbal, Painful responsive, unresponsive)	608(3.7%)	3,221(6.4%)
	Unconscious=No	Alert or verbal response based on the disability scale of AVPU	12529(76.1%)	45,915(90.6%)
	Missing data		3338(20.3%)	1,533(3.0%)

Table 4.2.4-2: Predictors used in Lowlaavar *et al.* model and their level of missingness both in derivation and validation datasets.

Predictor	Variable in model derivation dataset	Variable equivalent in external validation dataset	N in the Derivation datasets	N in the Validation datasets
Blantyre coma score	Abnormal Blantyre coma score (score <5)	Verbal response based on the disability scale of AVPU	Not provided	2023(2.3%)
		Missing data	Not provided	103(1.0%)

HIV diagnosis	Positive HIV diagnosis	Positive HIV diagnosis	66(5.1%)	850(1.0%)
Weight for age z-score (WAZ)	Severely stunted. (WAZ < -3)	Weight for age z-score < -3	206 (15.9)	481(4.5%)
	Underweight (WAZ < -2)	Weight for age z-score < -2	372 (28.6%)	2,649 (24.7%)
	Missing data		No provided	45(0.4%)
Mid-upper arm circumference	MUAC < 125mm	MUAC < 125mm	187 (14.5)	898(8.8%)
	MUAC < 115mm	MUAC < 115mm	94 (7.3)	292(2.8%)
	Missing data		No provided	531 (4.9%)

To avoid bias that may have resulted from excluding observations with missing data, we undertook multiple imputation to account for the uncertainty caused by missing data [114, 115]. To do this, we created 20 imputation datasets under the assumption of missing at random (MAR) mechanism. Variables to be imputed were ordered based on their levels of data missingness from low to high. This was meant to fully benefit from the chained equations of the imputing algorithm and to boost convergence. The simulation error in the multiple imputation was minimized by using 100 iterations between imputations. Validation of the prognostic model was carried out on each of the imputed dataset. Rubin's rules [116] were used to pool estimates from the 20 multiply-imputed datasets.

4.3 Results

4.3.1 Eligible population

The Clinical information Network's database had a total of 212,654 patients admitted and 162,329 patients that were eligible to be included in the validation cohort from all hospitals(n=20). We further applied model-specific exclusions as shown in Figure 4.2.2-1 to obtain n=50,669 and n=10,782 patients who were eligible for the external validation of RISC-

Malawi and Lowlaavar model respectively. Model performance results are presented using imputed dataset.

4.3.2 Characteristics of the cohort used in the external validation of the RISC-Malawi prognostic model.

We had n=50,669 patients who met the eligibility criteria to be included in the validation dataset of the RISC-Malawi model. Out of this cohort, pneumonia case fatality ratio was 8.7% which varied across hospitals ranging from 16.3% to 1.9%. Upon examining characteristics of these cohort, we observed that 3,221/50,669 (6.4%) of the patients were unconscious, of which 1,281/3,221 (39.8%) died. 3,042/50,669 (6%) of all patients were severely malnourished and 604 (19.9%) of them died. In addition, the data also suggested 8,949/50,669 (17.7%) patients experienced severe hypoxemia out of which 14% (1,253/8,949) of them died as shown in Table 4.3.2-1.

Table 4.3.2-1: Demographic and clinical characteristics of the cohort used to externally validate RISC-Malawi model.

	All patients	Survived	Died
Population	n=50669	46263/50669 (91.3%)	4406/50669 (8.7%)
Child-sex (Female)	22184/50669 (43.8%)	20001/22184 (90.2%)	2183/22184 (9.8%)
Age(months) Median (IQR)	13(7-24)	14(7-25)	9(6-16)
Moderate hypoxemia	3875/50669 (7.6%)	3591/3875 (92.7%)	284/3875 (7.3%)
Severe hypoxemia	8949/50669 (17.7%)	7696/8949 (86%)	1253/8949 (14%)
Moderately malnourished	8699/50669 (17.2%)	7988/8699 (91.8%)	711/8699 (8.2%)
Severely malnourished*	3042/50669 (6%)	2438/3042 (80.1%)	604/3042 (19.9%)
Wheeze present	6666/50669 (13.2%)	6181/6666 (92.7%)	485/6666 (7.3%)
Unconsciousness*	3221/50669 (6.4%)	1940/3221 (60.2%)	1281/3221 (39.8%)

	All patients	Survived	Died
Unconscious* defined as either Painful responsive or unresponsive in the disability scale of AVPU (Alert, Verbal, Painful responsive, unresponsive)			
Severe hypoxemia* defined as oxygen saturation <90%			
Severely malnourished* defined as Mid-upper Arm Circumference (MUAC) <11.5cm			
Moderately malnourished* defined as MUAC between 11.5cm and 13.5cm			

In general, a comparison of the patients' case-mix between the validation and derivation cohort of RISC-Malawi model suggested that characteristics were comparable between the two data sources however there was a higher prevalence of unconsciousness in the validation dataset (6.4%) than it was in derivation one (3.7%) as shown in Table 4.2.4-1.

4.3.3 Characteristics of the cohort used in the external validation of the Lowlaavar *et al.* 2016 models.

Since the derivation study of the Lowlaavar models included Blantyre coma score as a model predictor, the eligibility criteria to the model validation cohort included patients from 6 hospitals where the BCS data is collected. In this dataset 10,782 children met the eligibility criteria of which 570/10782 (5.3%) experienced in-hospital mortality. As defined in the model's derivation study, patients with a BCS<5 were considered to have abnormal BCS that was present in 1199 patients of whom 236(19.7%) died in hospital.

Table 4.3.3-1: Demographic and clinical characteristics patients from 6 hospitals who were included in the validation cohort of Lowlaavar *et al.* 2016 models.

Indicator	All patients	Survived	Died
Population	N=10782	10212/10782 (94.7%)	570/10782 (5.3%)
Gender (Female)	4508/10782 (41.8%)	4245/4508 (94.2%)	263/4508 (5.8%)
Age(months) Median (IQR)	24(14-42)	24(14-42)	22.5(11-38)
HIV diagnosis	75/10782 (0.7%)	63/75 (84%)	12/75 (16%)
Abnormal BCS*	1199/10782 (11.1%)	963/1199 (80.3%)	236/1199 (19.7%)
WAZ*	-0.5(-1.5-0)	-0.5(-1.5-0)	-1(-2-0)
MUAC*	14.3(13.5-15)	14.3(13.6-15)	14(13.1-14.8)
BCS*=Blantyre Coma Score WAZ*= Weight for Age Z-score MUAC*= Mid-upper Arm Circumference (MUAC) Abnormal BCS* defined as Blantyre coma core of <5			

4.3.4 Model performances in external validation dataset

The discriminatory ability (c-statistic) of the RISC-Malawi model was 0.77 (95% CI: 0.77 to 0.78) whereas the calibration slope was 1.04 (95% CI: 1.00 to 1.06), and calibration intercept was 0.81 (95% CI: 0.77 to 0.84) which is indicative of a poorly calibrated model since it is underestimating the risk (intercept >0) see Figure 4.3.4-1.

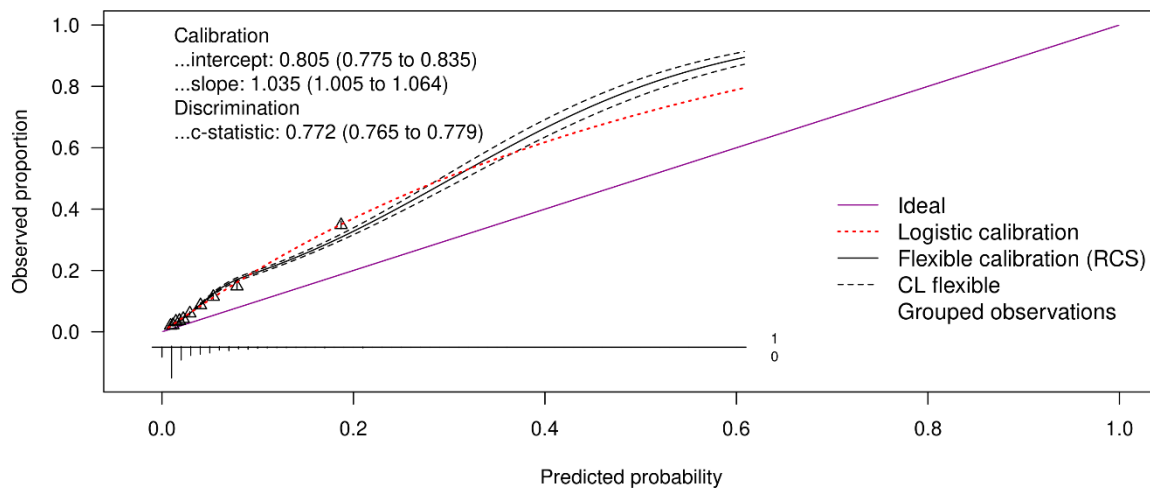


Figure 4.3.4-1: Performance of the RISC-Malawi model in an external validation dataset.

The figures show calibration curves and other model performance metrics. Key: RCS

denotes the Restricted Cubic Splines, and CL denotes the Confidence Limits (95%)

For the Lowlaavar *et al.* model, we computed the performance statistics for the 3 models and the findings were as follows; the primary model (model 1) which included 3 predictors (abnormal BCS, HIV+, weight for age z-score) had a c-statistic of 0.75 (95% CI: 0.72 to 0.77) while the calibration slope was 0.78 (95% CI: 0.71 to 0.84) and the calibration intercept was 0.37 (95% CI 0.28 to 0.46). The second model (model 2) had included the following 3 predictors: abnormal BCS, HIV+, and MUAC had a c-statistic of 0.78 (95% CI: 0.77 to 0.80) while the calibration slope was 0.82 (95% CI: 0.76 to 0.89) and the calibration intercept was 0.92 (95% CI 0.84 to 1.10). Lastly, the third model had 2 predictors namely abnormal BCS and MUAC. The model had a c-statistic of 0.71 (95% CI: 0.68 to 0.73) while the calibration slope was 0.73 (95% CI: 0.67 to 0.80) and the calibration intercept was 0.39 (95% CI 0.31 to 0.48) as shown in Figure 4.3.4-2.

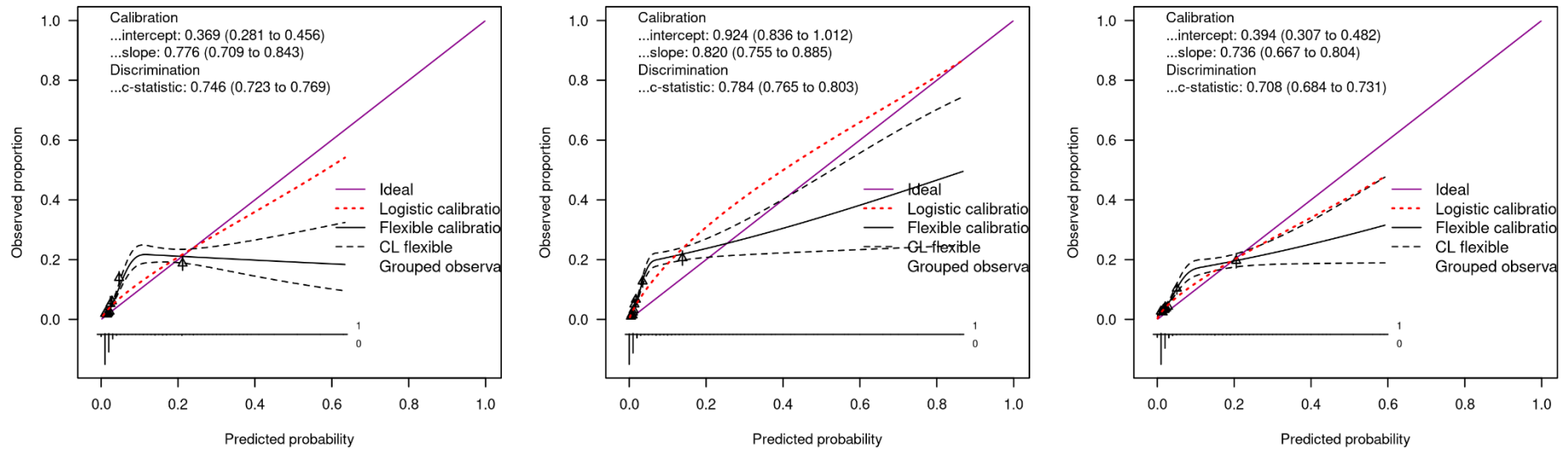


Figure 4.3.4-2: Performance of the Lowlaavar et al. 2016 models in an external validation dataset where abnormal Blantyre Coma Score (BCS) was defined as $BCS < 5$. The first panel to the left is the calibration curves of the primary model (Model I), the panel in middle are the calibration curves for model II, and the last panel to the right are the calibration curves of the model III. Key: RCS denotes the Restricted Cubic Splines, and CL denotes the Confidence Limits (95%).

4.3.5 Sensitivity Analyses

As a sensitivity analysis, we varied the criteria of determining pneumonia diagnosis in the validation cohort of RISC-Malawi such that instead of using danger signs (central cyanosis, grunting, indrawing, stridor, inability to drink, AVPU, and convulsion) as used in the original study to define pneumonia, we used clinical admission diagnosis of pneumonia. All other eligibility criteria remained unchanged, and this resulted to a sample size of 56,045 with a pneumonia case fatality rate of 7.6%. Evaluation of the RISC-Malawi model performance in the sensitivity analyses dataset suggested a reduced performance as compared to what was seen in the main analyses as shown in the Figure 4.3.5-1.

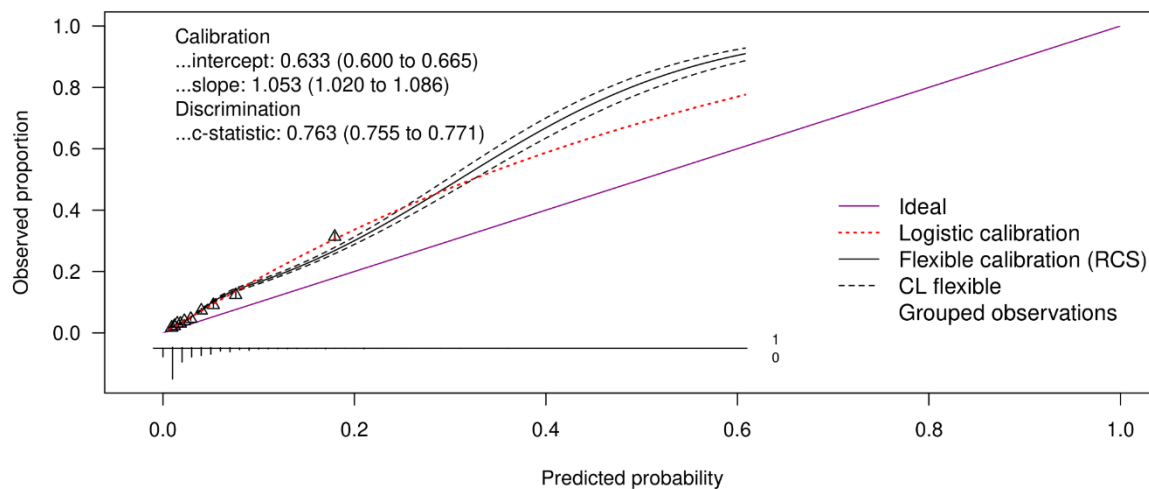


Figure 4.3.5-1: Performance of the RISC-Malawi model in a sensitivity analyses dataset (Pneumonia is defined based on the admission clinical diagnosis instead of danger signs). The values show calibration curves and other model performance metrics. Key: RCS denotes the Restricted Cubic Splines, and CL denotes the Confidence Limits (95%).

We also undertook a sensitivity analysis of the Lowlavaar models using patients from all hospitals(n=20) instead of the 6 hospitals as used in the main analyses. However, the abnormal BCS was defined using AVPU scores which is a disability scale such that patients who were at “V” during the clinical assessment by a physician were classified as having abnormal Blantyre scale. Those who met the eligibility criteria were 86,784 patients and in-hospital mortality was 4.7%(n=4045). Patients who were classified as having abnormal BCS were 2023(2.3%) out of which 268(13.5%) died. Additionally, there were 850 (1%) patients whose HIV status was known to be positive and 95(11.2%) of them died in hospital as shown in Table 4.3.5-1. Performance of the Lowlavaar models in the sensitivity analyses dataset were lower as compared to the main analyses as shown in Figure 4.3.5-2.

Table 4.3.5-1: Demographic and clinical characteristics of patients included in the sensitivity analyses dataset.

Indicator	All patients	Survived	Died
Population	N=86784	N=82739 (95.3%)	N=4045 (4.7%)
Gender (Female)	37683/86784 (43.4%)	35740/37683 (94.8%)	1943/37683 (5.2%)
Age(months) Median (IQR)	20(11-36)	20(12-36)	13(9-27)
HIV diagnosis	850/86784 (1%)	755/850 (88.8%)	95/850 (11.2%)
Abnormal BCS*	2023/86784 (2.3%)	1755/2023 (86.8%)	268/2023 (13.2%)
WAZ*	-1(-2-0)	-1(-2-0)	-1(-2-0)
MUAC*	14.2(13.4-15)	14.2(13.5-15)	13.5(12.5-14.5)
BCS*=Blantyre Coma Score WAZ*= Weight for Age Z-score MUAC*= Mid-upper Arm Circumference Abnormal BCS* defined as responding to Pain in the disability scale of AVPU (Alert, Verbal, Painful responsive, unresponsive)			

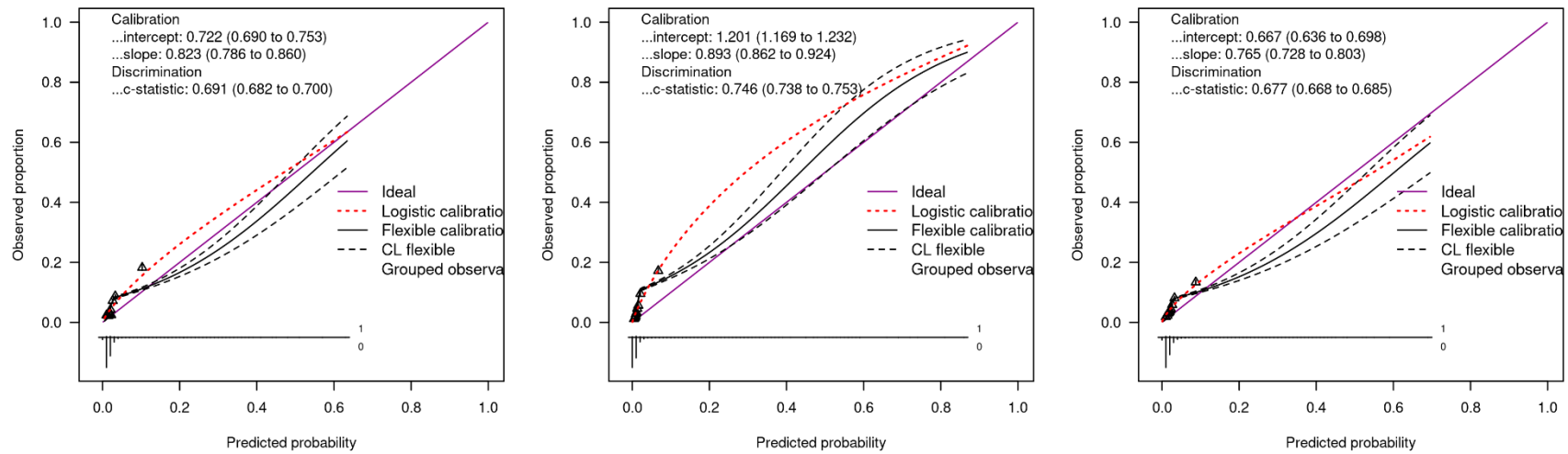


Figure 4.3.5-2: Performance of the Lowlaavar et al. 2016 models in an external validation dataset whereby Abnormal Blantyre coma score was defined using the disability scale of AVPU (Alert, Verbal response, Pain response, Unresponsive) such that patients who were not alert but responding to verbal stimuli were assumed to have abnormal Blantyre coma score. The first panel to the left is the calibration curves of the primary model (Model I), the panel in middle are the calibration curves for model II, and the last panel to the right are the calibration curves of the model III. Key: RCS denotes the Restricted Cubic Splines, and CL denotes the Confidence Limits (95%).

4.4 Discussion

4.4.1 Summary of key findings

Validation of existing prognostic models to identify children at risk of deterioration in diverse settings is the first step towards wider clinical application of clinical prediction rules. In this chapter, 4 prognostic models were externally validated that were originally designed by two studies to identify children at an increased risk of in-hospital mortality in low resource setting [41, 108]. Using a diverse population of children admitted to 20 hospitals from 2014 to Dec 2021, we performed areas under the curve analysis to assess the discriminatory ability as well as the calibration levels of these four prognostic scores. All models had fair discriminatory values (AUC 0.70-0.79) however, all of them markedly underestimated the mortality (calibration intercept > 0). This leads to misclassification of patients who are at an increased risk of deterioration. The model performance measures were even lower when these models were validated using the sensitivity analyses datasets where we varied the definitions of abnormal Blantyre Coma score, and pneumonia from how it was defined in the original study by Lowlavaar *et al.*, and Hooli *et al.* for RISC-Malawi respectively. This demonstrates the value addition of defining predictors as used in model derivation study.

The sub-optimal performance of these models in the CIN datasets may be due to having more diverse patient populations, and different case-mix. Although we attempted to make sure that the patients characteristics in validation and derivation cohort were as similar as possible, we observed that the cohort we used to validate the RISC-Malawi model had a higher childhood mortality rate (8.5%) than the original patient group (3.2%). Despite this difference the discriminatory ability of the RISC-Malawi in the validation cohort had an AUC 0.77(95% CI: 0.77 to 0.78) which was nearly alike to that observed in the model derivation cohort 0.79 (95% CI: 0.76 to 0.82).

While the calibration intercepts of all models we externally validated suggested underestimation of the risk of mortality in their predictions, calibration slopes of the same models illustrated that these predictions were too extreme especially for the Lowlaavar *et al.* models whose calibration slopes were all <1 . On the other hand, predictions of the risk of pneumonia-related mortality by RISC-Malawi model were too low as judged by the calibration slope of >1 . The result may be partly explained by the inclusion of more physical examination variables as prognostic factors which could potentially make a model to underperform in an external dataset because of the variations in inter-observer agreement which is more common in physical examination findings [117, 118]. It is encouraged to include prognostic factors that do not have inter-observer variations such as blood lactate and other biomarkers including C-reactive protein, procalcitonin, etc., in the settings where these tests are available. However, while such biomarkers might be having better prognostic values and hence attractive to be included in the prognostic models, they may not be readily available in limited-resource settings and are costly to undertake.

In literature, RISC-Malawi model has been subjected to an external validation in a diverse cohort of hospitalized children from the World Health Organization's study group whose study patients were pooled from 10 studies on pediatric pneumonia from different countries [119]. In this cohort there were 17,864 who met the eligibility criteria with a pneumonia case fatality ratio of 4.9%. The RISC-Malawi score in that validation study had fair discriminatory value (AUC = 0.75, 95% CI = 0.74-0.77) which was not very different from what was obtained in our validation study even though our validation cohort had a higher pneumonia case fatality ratio of 8.7% (Table 2). Furthermore, we could not determine RISC-Malawi's calibration measures in the validation study since these were not reported. To our knowledge, Lowlaavar *et al.* models have not been externally validated in any setting.

4.4.2 Limitations

While CIN database is a rich source of data that is routinely collected from several hospitals over a period and hence suitable for model development and validation, by design these data were not meant for such purposes instead it was an essential initial step in efforts to understand and improve care in Kenyan hospitals. This led to missing data in variables of interest for many children which resulted to multiply imputing the data a task that was computationally prohibitive. However, CIN dataset had a substantial effective sample size required for external validation studies [120]. Lastly, even though we attempted to make the validation population as similar as possible to that used in the derivation of the models we externally validated, we didn't exclude children who carried more than one diagnosis concomitantly, which could explain the reason why the validation case mortality rate being twice to that of the derivation cohort.

4.4.3 Fulfilled knowledge gaps and what to be done next.

In the literature it is more common to find model development studies than validation ones. Hence there are a lot of models which are running a risk of not being utilized in the clinical practice because they have not been externally validated using diverse population as expected hence become wasted research efforts. When evaluating a model for risk stratification, researchers should utilize pre-existing knowledge and, if available, validate and update an existing model within a similar setting instead of building a new model from scratch with all the drawbacks of overfitting and lack of reproducibility. In this study we have subjected 4 models to an external validation study to determine their clinical utility. In the ideal case of perfect validity where scores have $AUC \geq 0.8$, calibration intercept = 0, and calibration slope = 1 then the model could be recommended for use in the clinical application. However, if the model deviate from the ideal case, then there is evidence of miscalibration and model recalibration should be performed [92, 121]. Our findings have suggested that the 4 models have significant

miscalibration and hence underlining the necessity of recalibration as a next step which is reserved for Chapter 5.

4.5 Conclusions

Despite the common challenge of prognostic models showing reduced performance when applied to populations different from the one used for their development, none of the externally validated models in this chapter demonstrated exceptional discrimination, with an $AUC \geq 0.8$, while maintaining precise risk estimation based on calibration statistics. Consequently, using these models in settings other than those in which they were originally developed may not be advisable with a high degree of confidence. Our findings suggest that recalibrating these models or considering the creation of new prognostic models with improved sensitivity and specificity for identifying children at risk of in-hospital mortality may be a worthwhile pursuit

Chapter 5

Recalibrating Prognostic Models

5.1 Introduction

Prognostic models predict patients' risk of deterioration or poor outcome and good models can inform clinical treatment or follow-up plans [25] but developing new models without investigating performance of existing models wastes potentially important historical data and research efforts [72]. External validation of published prognostic models in comparable settings or populations is recommended in establishing model transportability and generalizability [122-124].

Most clinical prediction models may not perform well in external validation and end up being rejected because of poor predictive performance. This is partly because clinical environments continuously evolve in various ways, including shifts in clinical practice, even though clinical practice guidelines tend to standardise this [125]. Other reasons include change of patient's management such as use of aggressive treatment therapies e.g. use of higher molecules of antibiotics as opposed to the first-line, and introduction of new vaccines e.g., RTS,S/AS01 which is a world's first malaria vaccine [113]. Such interventions may change the prevalence and clinical presentations of common childhood illnesses, and thus, a clinical prediction model developed before these interventions would perform poorly when validated in such settings. Variation in case-mix, different time points of model development and validation, and dataset drift also contribute towards deterioration of the model performance when applied in new samples hence a need for model recalibration to contextualize to the local settings [126].

Model updating is suggested once there is evidence of poor model calibration but acceptable discriminatory abilities in an external validation study [91, 92].

In chapter 4, the four identified prognostic models were externally validated but their calibration estimates suggested an underestimation of in-hospital paediatric mortality risk.

In this chapter, we aim to recalibrate these models (Respiratory Index of Severity in Children (RISC-Malawi) [41] and three other models developed by Lowlaavar *et al.* [108]) using regression coefficients updating strategies and determine how much their performances improve.

The rest of this chapter is structured as follows; section 5.2 provides details of the model calibration metrics, assembling if the recalibration cohort, sample size determination for the model recalibration, and recalibration strategies. Section 5.3 provides model recalibration results; the chapter concludes with discussion and the conclusion sections in section 5.4 and 5.5 respectively.

5.2 Methods

5.2.1 Models' calibration metrics

The threshold for a perfectly calibrated score is a model with a calibration slope of 1 and calibration intercept (calibration-in-the large) of 0 or an identity line of 45° in the calibration plot indicating limited chances of over/underestimating the risk of bad outcomes when used in clinical practice. Although it is not clear how close these metrics should be to the set thresholds for the model to be acceptable, there is consensus from the literature that a model has good calibration if the intercept is close to 0 and the slope is close to 1 [127]. For instance, a model slope of 0.95 was termed “good calibration” by Philips *et al.*, [128] and Nakhjavan *et al.* [129] termed a model with a slope of 0.97 and an intercept of 0.006 “proper calibration”.

5.2.2 Details of models to be recalibrated.

The RISC-Malawi [41] model and 3 models by Lowlaavar *et al.* 2016 [108] were identified in chapter 3 and externally validated in chapter 4. In brief, RISC-Malawi is a Respiratory Index of Severity in Children (RISC) that was developed using prospectively collected clinical data

from a cohort of 14,665 hospitalized children aged 2-59 months with pneumonia in Malawi between 2011–2014. The three models by Lowlaavar *et al.* 2016 [108] were developed utilizing a two-site prospective observational study in Uganda which enrolled 1307 children between 6 months and 5 years admitted with a proven or suspected infection. In chapter 4, findings suggested that while they had fair discriminatory ability (c-statistics of 0.70 to 0.79 [86, 87]), they were poorly calibrated as judged from their calibration slopes and intercepts of these models as shown in Figure 5.2.2-1. For instance, RISC-Malawi had a c-statistic of 0.77 (95% confidence interval (CI): 0.77 to 0.78), a calibration slope of 1.04 (95% CI: 1.00 to 1.06), and the calibration intercept was 0.81 (95% CI: 0.77 to 0.84). Lowlaavar *et al.* 2016 had a c-statistic of 0.75 (95% CI: 0.72 to 0.77), calibration slope was 0.78 (95% CI: 0.71 to 0.84), and the calibration intercept was 0.37 (95% CI 0.28 to 0.46).

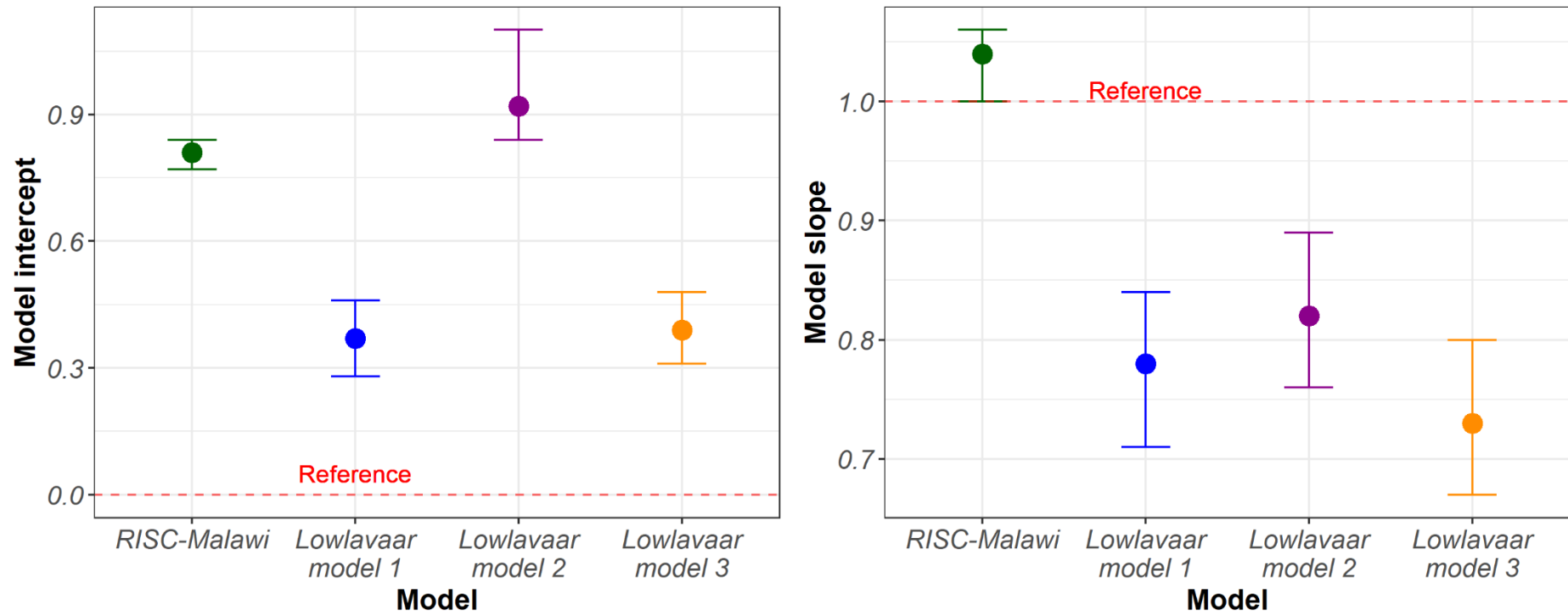


Figure 5.2.2-1: Model intercept and model slope of the four models suggesting that models were not well calibrated. These estimates were obtained from external validation study

5.2.3 Availability of model predictors in the recalibration cohort

For RISC-Malawi model, all predictors were available across all 20 hospitals contributing to model updating dataset except for the predictor called *unconsciousness* which was unavailable. However, we recoded this predictor based on the disability scale of AVPU (Alert, Verbal response, response to Pain, Unresponsive) such that a patient was assumed to be unconscious if the clinician rated them as either “P” (responding to pain only) or “U” (unresponsive). AVPU is known for the assessment of the patient’s brain function hence used for the determination of the level of consciousness[130]. For the Lowlaavar model, all predictors were available in all hospitals except for the *Blantyre Coma Score* which was available in only 6 hospitals as from September 2019. Therefore, for the updating and testing of the Lowlaavar models we only used data from the 6 hospitals for patients admitted as from September 2019 through December 2021 when use and recording of Blantyre Coma Score was introduced in the six hospitals.

5.2.4 Eligibility criteria for model recalibration cohort

To determine the appropriate patients to be included in the cohort of model recalibration, we applied eligibility criteria as was used in the original model derivation studies [41, 108]. In summary, for the RISC-Malawi model we included children aged 2-59 months with admission diagnosis of pneumonia defined as either cough or difficult breathing, and any of the danger signs namely central cyanosis, grunting, chest wall indrawing, stridor, inability to drink/breastfeed, convulsing or not alert based on the disability scale of AVPU scale. For the Lowlaavar models, we included children aged 6-60 months admitted with any confirmed or suspected infectious diseases. To achieve this eligibility criteria, we filtered out all patients with non-communicable diseases. In each of the two model recalibration cohorts, we excluded children admitted for surgery or with burns, trauma, road traffic accidents, those with poisoning

such as organophosphate ingestion, and those patients admitted during healthcare workers' strike.

For each model's eligible dataset, we split the data into model updating dataset (for recalibrating the model) and testing dataset (for assessing model performance after updating). For RISC-Malawi model, 50,669 patients met the eligibility criteria, the updating dataset included 30,343 patients admitted across all 20 hospitals from January 2014 through December 2018 while the model testing dataset included 20,326 patients admitted in the same hospitals from January 2019 through December 2021 as shown in Figure 5.2.5-1. For Lowlaavar models there were 10,782 patients who met the eligibility criteria in 6 out of 20 hospitals. 7521 of these patients admitted in 4 hospitals were used to update the models and the remaining 3261 patients from 2 hospitals were used to test these models.

5.2.5 Sample size for model recalibration

Following sample size calculation approaches by Riley *et al.* [131] that took into account the c-statistics of the original models, the number of parameters in the original model, and the prevalence of the outcome (in-hospital mortality) in the derivation cohort, we computed the sample sizes required to recalibrate each of the four models assuming an acceptable difference of 0.05 between the apparent and adjusted *R-squared* of the original model. Minimum sample sizes required for each of the 4 models are provided in Table 5.2.5-1. For example, while sample size calculation approaches required a minimum sample size of 1619 for RISC-Malawi model our model updating and testing datasets exceeded this having sample sizes of 30,343 and 20,326 respectively as shown in Figure 5.2.5-1. In addition, with mortalities of >1000 in RISC-Malawi models datasets, and >200 in Lowlaavar models' datasets, the events-per-variable ratios exceeded the recommended ratio of 20, given the >100 events per variable for each model [131, 132].

Table 5.2.5-1: Minimum required sample sizes for recalibration of identified models.

Model	c-statistic in the derivation cohort	Number of parameters in the original model	Outcome prevalence in the derivation cohort	Margin of error in estimation of intercept (assumption)	Difference between apparent and adjusted R-squared	Minimum required sample size
RISC-Malawi model	0.79	7	3.2%	5%	5%	1619
Lowlaavar model 1	0.85	3	5%	5%	5%	285
Lowlaavar model 2	0.84	3	5%	5%	5%	307
Lowlaavar model 3	0.82	2	5%	5%	5%	239

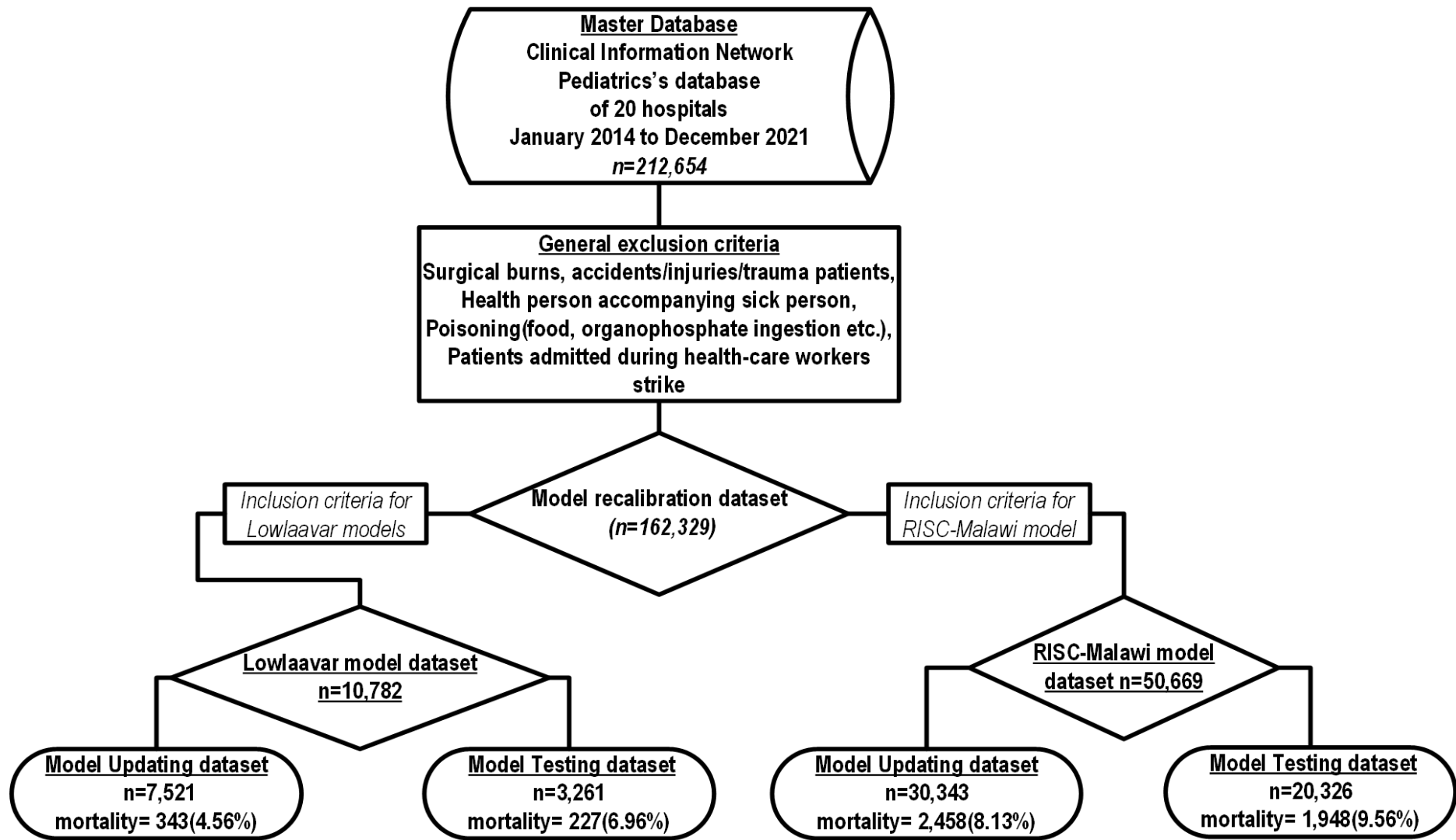


Figure 5.2.5-1: Populations used to update and test RISC-Malawi model and 3 models by Lowlaavar et al. 2016

5.2.6 Assessment of missing data in the model recalibration cohort

Model recalibration entails numerical adjustment of the model's intercept and regression coefficients of the prognostic factors by a common numerical value. In such computations all prognostic factors are expected to have data for each patient in a cohort, otherwise records with incomplete data are deleted from the analysis resulting to "complete case analysis" that could lead to loss of statistical power and potentially yield biased estimates[133]. Missing data assessment suggested that in the cohort for updating RISC-Malawi model 68.3% of the patients' records risked being dropped from the analysis because of the incomplete data in the required variables while 5.2% of the records in Lowlaavar models' cohort would also be discarded through a complete case analysis.

5.2.7 Model recalibration strategy

In the model recalibration strategies, we employed equations as described in detail in chapter 3 section 3.5.1 but the following are the equations used in the model recalibration.

Recalibration-in-the-large equation

$$\log\left(\frac{P(\text{in-hospital mortality})}{1 - P(\text{in-hospital mortality})}\right) = (\alpha + \alpha_{\text{correction factor}}) + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n$$

Logistic calibration equation

$$\log\left(\frac{P(\text{in-hospital mortality})}{1 - P(\text{in-hospital mortality})}\right) = (\alpha_{\text{correction factor}}) + (LP_{\text{original}} \times \beta_{\text{correction factor}})$$

5.2.8 Assessing performance of the recalibrated prognostic model in the testing dataset.

For each model we separately applied the two recalibration strategies (intercept only, and intercept and slope) as described above. Based on the adjusted model, we computed linear predictor for each patient in the model-specific testing dataset (see Figure 5.2.5-1) which was used to compute patient's predicted risk of mortality via a logistic function. Model performance was determined using two metrics namely discriminatory index and model calibration. Discriminatory ability was determined using the c-statistic (value 0–1, discriminative if > 0.7) [84, 85] while the calibration was measured using the calibration slope that summarises agreement between predicted and observed risks and it ranges from 0 to 1 with values near 1 showing better accuracy while values < 1 suggesting predicted risks that are too extreme, and calibration intercept which indicates the extent that predictions are systematically too low or too high, with predicted risks under-estimated if > 0 or over-estimated if < 0 [88].

5.3 Results

5.3.1 Characteristics of the model recalibrating cohorts.

The eligibility criteria for RISC-Malawi model were met in 50,669 patients from all 20 hospitals which were split into model recalibrating ($n=30,343$) and testing ($n=20,326$) datasets. The distribution of patient characteristics in recalibrating and test datasets were similar, although the test set had slightly higher mortality 1948 (9.6%) than the updating dataset 2458 (8.1%). This finding was not unexpected because in the cohort for model testing, cases of severe hypoxemia were 24% which were almost twice that of model updating (13.4%) as shown in Table 5.3.1-1. However, we noted that cases of severe hypoxemia in the RISC-Malawi's original study was 12.7% which was comparable with that of model updating dataset.

Table 5.3.1-1: Distribution of clinical characteristics of the cohort used to recalibrate and test RISC-Malawi model.

	Updating dataset (N=30343)	Testing dataset (N=20326)	All patients (N=50669)
Mortality	2458 (8.1%)	1948 (9.6%)	4406 (8.7%)
Child-sex (Female)	13380 (44.1%)	8804 (43.3%)	22184 (43.8%)
Age in months Median [Min, Max]	13.0 [2.00, 59.0]	13.0 [2.00, 59.0]	13.0 [2.00, 59.0]
Moderate hypoxemia*	1971 (6.5%)	1904 (9.4%)	3875 (7.6%)
Severe Hypoxemia*	4071 (13.4%)	4878 (24.0%)	8949 (17.7%)
Moderately malnourished*	5245 (17.3%)	3454 (17.0%)	8699 (17.2%)
Severely malnourished*	1882 (6.2%)	1160 (5.7%)	3042 (6.0%)
Wheezing	3837 (12.6%)	2829 (13.9%)	6666 (13.2%)
Unconscious*	1774 (5.8%)	1447 (7.1%)	3221 (6.4%)
Unconscious* defined as either Painful responsive or unresponsive in the disability scale of AVPU (Alert, Verbal, Painful responsive, unresponsive) Moderate hypoxemia* defined as oxygen saturation 90%-92%% Severe hypoxemia* defined as oxygen saturation <90% Severely malnourished* defined as Mid-upper Arm Circumference (MUAC) <11.5cm Moderately malnourished* defined as MUAC between 11.5cm and 13.5cm			

For the Lowlaavar models, there were 10,782 patients meeting the eligibility criteria in 6 out of the 20 hospitals with an overall in-hospital mortality of 5.3%. A sub-analysis to understand the distribution of mortality in the cohort revealed that mortality was higher in the testing dataset 227(7.0%) as compared to the updating dataset 343 (4.6%). In addition, patients classified to have abnormal Blantyre Coma Score (n=1096,10.2%) had a higher mortality of 19.4%. On further examination of the abnormal BCS population, we noted that patients in the testing cohort had a relatively higher mortality of 23.0% as compared to those in updating cohort 17.9%. The distribution of other model predictors was similar between testing and updating datasets as shown in Table 5.3.1-2.

Table 5.3.1-2: Demographic and clinical characteristics of the cohort used to recalibrate and test Lowlaavar model.

	Updating (N=7521)	Testing (N=3261)	All patients (N=10782)
Mortality	343 (4.6%)	227 (7.0%)	570 (5.3%)
Child-sex (Female)	3131 (41.6%)	1377 (42.2%)	4508 (41.8%)
Age in months Median [Min, Max]	24.0 [6.00, 60.0]	24.0 [6.00, 60.0]	24.0 [6.00, 60.0]
HIV diagnosis	16 (0.2%)	59 (1.8%)	75 (0.7%)
Abnormal Blantyre Coma Score	761 (10.1%)	335 (10.3%)	1096 (10.2%)
Weight for Age Z- score	-0.500 [-4.00, 4.00]	-1.00 [-4.00, 4.00]	-0.500 [-4.00, 4.00]
Mid-upper Arm Circumference (MUAC) in centimeter	14.2 [7.00, 21.0]	14.3 [8.20, 21.7]	14.3 [7.00, 21.7]

5.3.2 Predictive performance of the recalibrated RISC-Malawi model

To adjust the predictive performance of RISC-Malawi, we used two methods (intercept only, and logistic recalibration) as earlier described. The original model slope was 1.04 (95% CI: 1.00 to 1.06) indicating regression coefficients were slightly small (close to zero) and thus underestimating in-hospital mortality predictions in the new patients. On the other hand, the calibration intercept was 0.81 (95% CI: 0.77 to 0.84) indicating that the predicted probabilities are systematically too low. To adjust the RISC-Malawi model we used the correction factor for slope $\alpha_{correction\ factor} = 0.84$ and for intercept $\beta_{correction\ factor} = 0.326$ which were estimated from logistic recalibration model using the updating dataset and equation (3.4.3-1) was used in model adjustment. The adjusted model showed an improvement in model intercept by 0.04 (95% CI: -0.001 to 0.08) compared to the original one. However, upon assessing the same model in a separate dataset (testing), the model intercept deteriorated slightly to 0.12 (95% CI: 0.07 to 0.17), and model slope also dropped from 1.33 (95% CI: 1.28 to 1.38) in the model updating dataset to 1.08 (95% CI: 1.03 to 1.13) in the model testing dataset as

summarised in Figure 5.3.2-1. As compared with the derivation cohort, the discriminative ability of the RISC-Malawi was not any different in the updating dataset c-statistic 0.78 (95% CI: 0.78 to 0.79) but this was lower in the testing dataset 0.75 (95% CI: 0.74 to 0.76) as shown in Figure 5.3.2-3. Results of the intercept only method improved model intercept, but it suggested slope model adjustment was required as provided in the figure.

Table 5.3.2-1: Correction factors for model intercept and model slope

Model	Intercept correction factor(α)	Slope correction factor (β)
Lowlaavar model 1	-0.469	0.757
Lowlaavar model 2	-0.207	0.699
Lowlaavar model 3	-0.450	0.763
RISC-Malawi	0.326	0.846

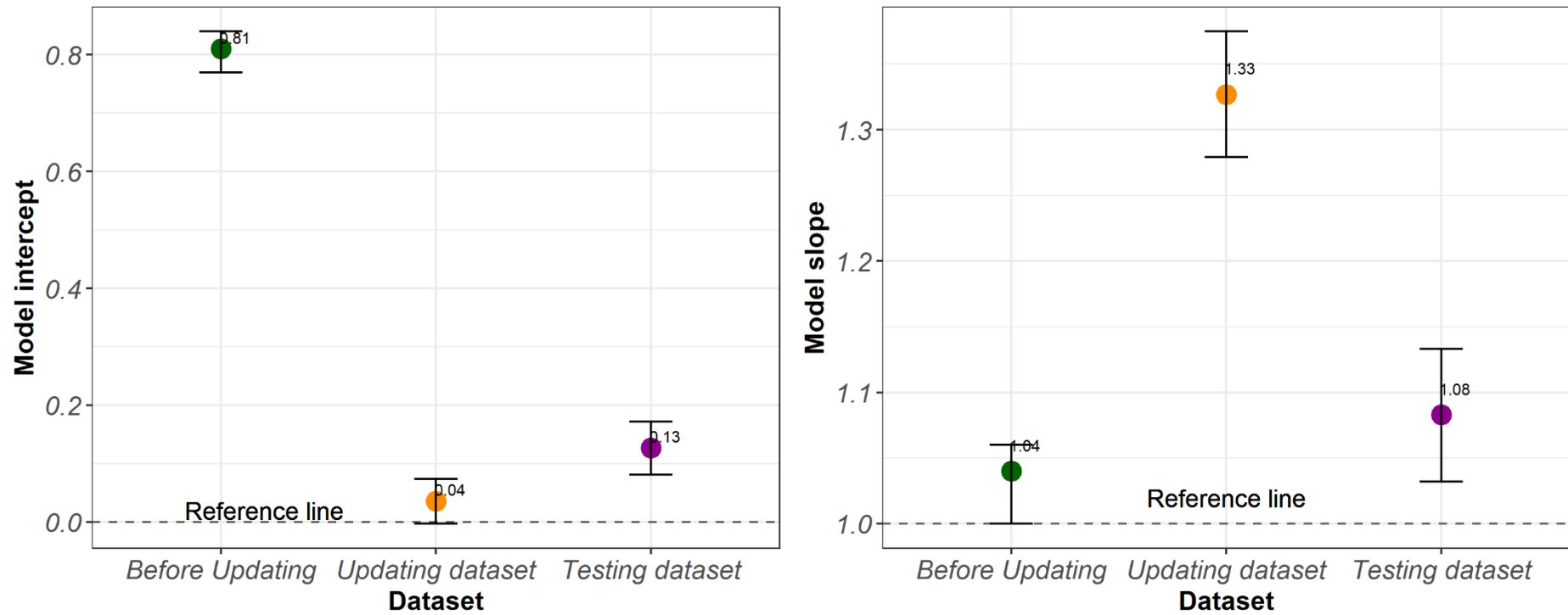


Figure 5.3.2-1: RISC-Malawi model calibration performance in various datasets. The figure in the left show calibration intercept while that on the right shows model slope. The coloured points and the 95% confidence intervals (shown as errors bars) shows the model calibration performances in the external validation, updating dataset (for model recalibration), and in the testing dataset. The dotted line denotes the references of the model intercept ($\alpha=0$) and slope ($\beta=1$) for a perfect calibrated model.

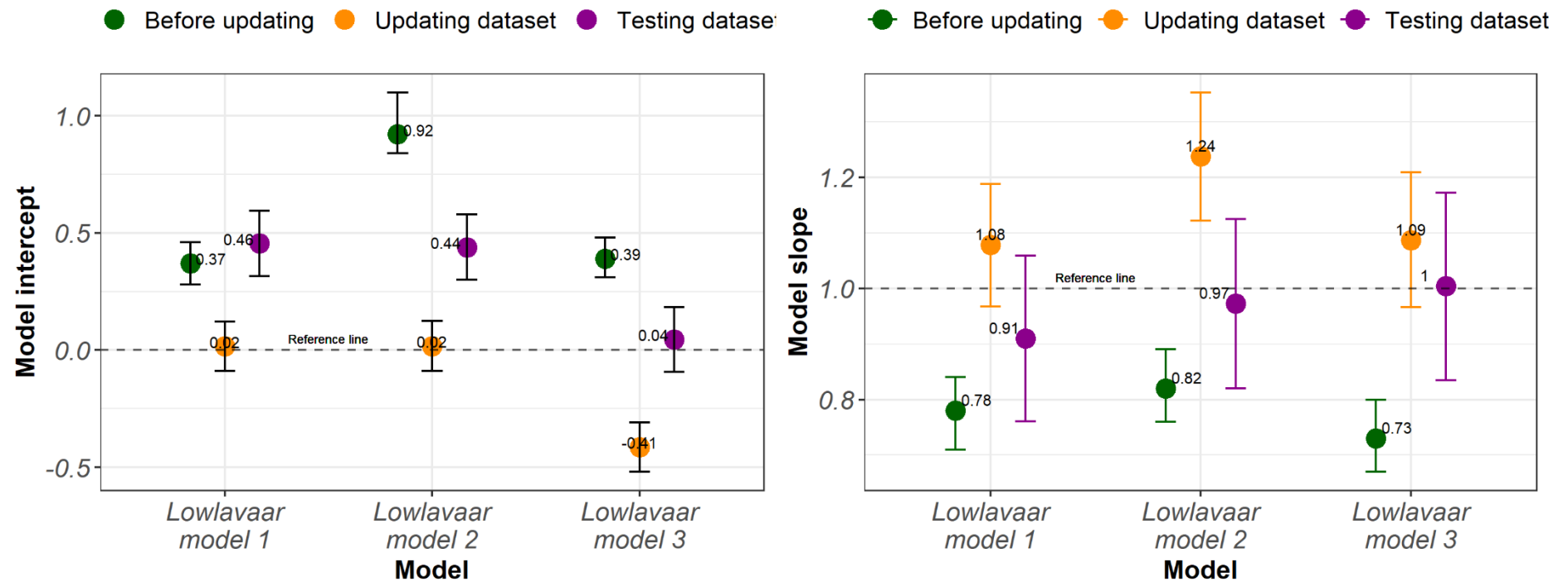


Figure 5.3.2-2: Calibration performance of Lowlaavar models in various datasets. The figure in the left show calibration intercept while that on the right shows model slope. The coloured points and the 95% confidence intervals (shown as errors bars) shows the model calibration performances in the external validation, updating dataset (for model recalibration), and in the testing dataset. The dotted line denotes the references of the model intercept ($\alpha=0$) and slope ($\beta=1$) for a perfect calibrated model.

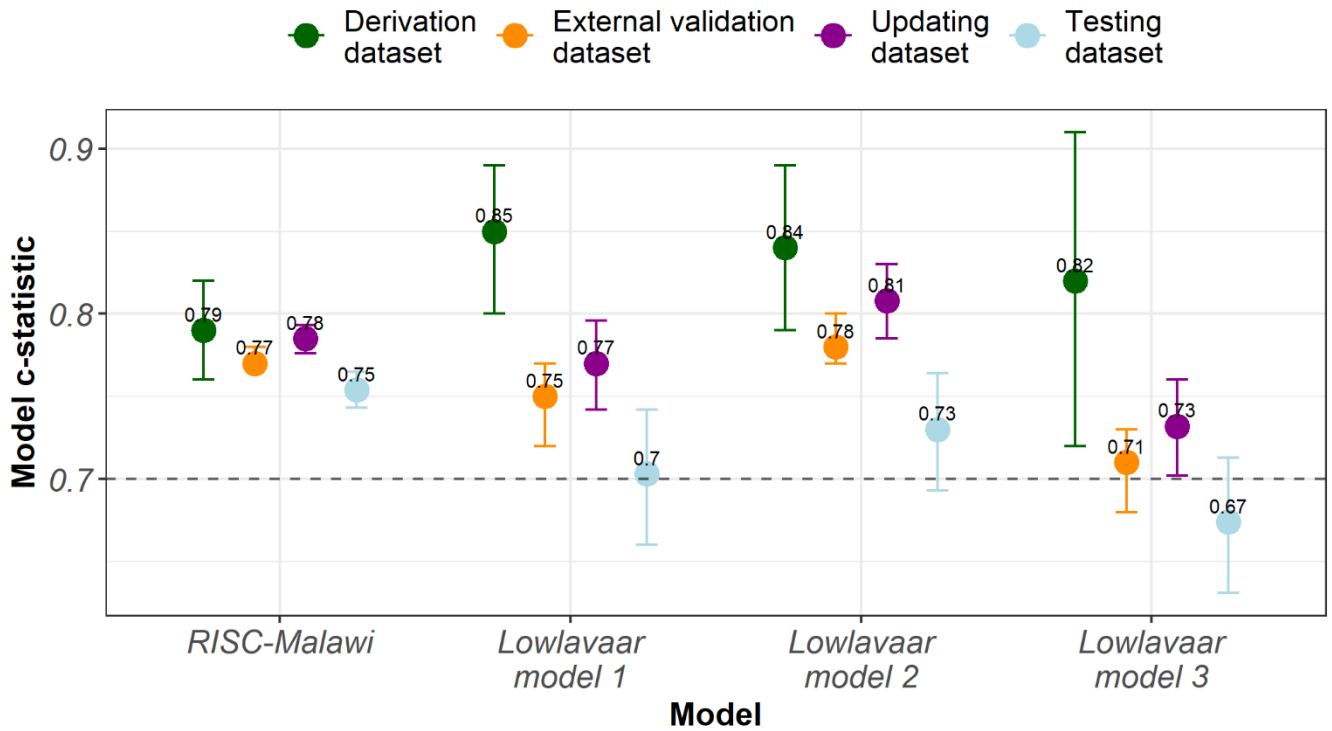


Figure 5.3.2-3: Discriminatory ability of the four models (RISC-Malawi, and the 3 models by Lowlavaar et al.) in various datasets. The coloured points and the 95% confidence intervals (shown as errors bars) shows the c-statistics of the in the derivation dataset, external validation, updating (for model recalibration), and in the testing dataset. The dotted line denotes a fair discriminatory ability of the model (Area Under Curve of 0.7)

5.4 Discussion

In this chapter we sought to recalibrate the four models which were externally validated win chapter and found to be over/underestimating the predicted risk of the in-hospital mortality. In general, the performance of a model with promising discriminatory ability (AUC >0.7) but with poor calibration can be improved using recalibration strategies. To achieve the objective, we used a relatively large sample sizes powered enough to update and test these models as shown in Figure 5.2.5-1. In addition, data used in this chapter has both temporal and spatial

richness since it has been collected from 20 county referral hospitals from 2014 through 2021. We explored both calibration-in-the-large adjustment and logistic calibration as recalibration strategies. Comparing results of the two model updating strategies, we observed that logistic recalibration was effective as expected than the recalibration-in-the-large method because the latter only adjusts the average predicted risk. The findings of model updating suggest that while the calibration of the models improved after recalibrating in updating dataset and upon testing in the test dataset, the differences in calibration performances in before and after updating were small as shown in Figure 5.3.2-1 for RISC-Malawi and Figure 5.3.2-2 for Lowlaavar models. However, these differences may not be clinically meaningful in practice since it has not met the required thresholds. The threshold for a perfectly calibrated score is a model with a calibration slope of 1 and calibration-in-the large of 0, or an identity line of 45° in the calibration plot indicative of limited chances of over/underestimating the risk of bad outcomes when used in clinical practice. The model 3 of Lowlaavar et al. appeared to have met the calibration thresholds however its discriminatory ability was below the minimum acceptable threshold of $AUC > 0.7$. Since the objective of this chapter was not to refit models, the recalibration strategies employed here do not change the ranking of the patients' predicted risk of in-hospital mortality, and as a result do not affect models' discriminatory ability. It is possible that a drop of AUC in the test dataset could be due to the unfortunate split sampling between updating and testing datasets or could indeed be due to chance. Based on this understanding, Lowlaavar model 3's low AUC in the testing dataset underscores a need to validate published prognostic models across plausibly similar contexts to ascertain if their discriminatory ability is consistent as expected. In summary, models we recalibrated exhibited poor performance even after adjustment to the local context. It is noteworthy that a miscalibrated prognostic model has been termed to be "clinically harmful" since it reduces the net benefit of its applicability in identifying risky patients for treatment [134].

Performance of the updated models can be explained by predictor-outcome associations being substantially different population in derivation, updating and in testing dataset [91]. For instance, as compared with pneumonia case-fatality in the derivation dataset which was 3.2%, the dataset used to update and test RISC-Malawi model had a higher pneumonia case-fatality of 8.13% and 9.56% respectively as shown in the Figure 5.2.5-1. On the other hand, mortality in the dataset used to adjust Lowlaavar models were only slightly higher than that of derivation cohort.

Models often exhibit relatively poor predictive performance in the external validation studies. This is partly because clinical environments continuously evolve in various ways including shift in clinical practice even though clinical practice guidelines tend to standardize this [125]. Other reasons include change of patient's management such as use of aggressive treatment therapies e.g. use of higher molecules of antibiotics as opposed to the first-line, and introduction of new vaccines e.g., RTS,S/AS01 which is a world's first malaria vaccine [113]. Such interventions may change prevalence and clinical presentations of common childhood illnesses and thus would make a clinical prediction model developed before these interventions perform poorly when validated in such settings. Variation in case-mix, different time points of model development and validation, and dataset drift also contribute towards deterioration of the model performance when applied in new samples [126] hence a need to adjust model to contextualize to the local settings. To do this, that's where the recalibration of the model intercept helps to capture nuances brought about by the variations in settings which are hard to be incorporated in the model predictors.

While it is more common for researchers to develop new prognostic models and sometimes even without a regard to methodological rigour [122, 123], there is a growing interest among researchers to recalibrate existing models to align with local context and be applied in clinical practice if found to be suitable. However, in the literature of prognostic research, it has not been established the acceptable differences between the expected calibration thresholds and the observed model calibration performances, it is also not clear on the number of external validations a prognostic

model is expected to have been subjected to before model updating is justified. In addition, even if a predictive model would be subjected to repeated model recalibrations, it is likely that prediction performance will plateau where no further meaningful gain will be realised [135]. It is therefore important for researchers to consider ensemble machine learning techniques such as stacking which are useful in combining predictive abilities of various competing models to yield a meta-model whose predictive performance would certainly be relatively better than that of a single model [136].

5.5 Conclusion

Due to inherent sampling variations, the performance of any model may exhibit slight discrepancies when applied to new patient samples. The conventional approach involves creating entirely new models, often resulting in a loss of valuable insights from prior prognostic modeling endeavors and an increased risk of overfitting, thereby diminishing the models' generalizability. In this chapter, we sought to enhance the performance of existing prognostic models by leveraging methodological strategies applied to large datasets. Our exploration led to the discovery that these models could be effectively improved through straightforward recalibration techniques, albeit not meeting the anticipated calibration thresholds of 0 for the model intercept and 1 for the model slope. This underscores the need for a computational approach that amalgamates these models into a meta-model, thereby enhancing their out-of-sample predictive performance which is devoted for next chapter of this report.

Chapter 6

Accounting for model uncertainty through stacking of predictive distributions of prognostic models

6.1 Introduction

In developing predictive models, researchers/statisticians often consider a wide range of competing models as a potential representation of the observed data with the assumption that the true data-generating model exists among the considered models. A single model is then selected from the list of competing models based on probabilistic or resampling criteria. Probabilistic model selection methods include AIC (Akaike Information Criterion) and BIC (Bayesian Information Criterion). On the other hand, the resampling methods include bootstrap, random train/test split of the data, cross-validation such as K-fold methods, or leave-one-out. Based on the output of the "best" chosen model, statistical inferences and conclusions are drawn with an implicit assumption that the distribution function of the selected model constitutes the actual data-generating model (DGM). This approach is commonplace in literature since it is standard statistical practice. However, it is not entirely satisfactory since it ignores model uncertainty in the selection of a suitable probability distribution function that might result in a misleading statistical inference and over/underestimation of the risk of the outcome of interest [20, 21]. Traditional methods of developing predictive models typically overlook model uncertainty hence yielding a model

with a tendency to overstate the goodness-of-fit between model and data, causing the model to lose predictive power when applied to independent datasets [137, 138].

Aware that prognostic models are to be used to predict in-hospital mortality and be used by clinicians in hospitals as a 'job aid' to identify children at an increased risk of deterioration, every little bit of model performance matters substantially.

In this chapter, given a forementioned background, a novel approach is proposed that leverages on Stacking of Bayesian predictive distributions of the candidate models. This machine learning technique is analogous to the 'wisdom of the crowd' phenomenon [139], which arguably outperforms any form of single model selection techniques in terms of out-of-sample predictive performances [140-142]. This is achieved by fitting each model to the data using Bayesian inference and calculating the posterior predictive distributions for each model. These distributions represent the range of predictions the models would make for new data. The next step involves determining the weights for each model. To achieve this, a loss function is used such as the Kullback-Leibler divergence, which quantifies the difference between the predicted distributions and the true distribution [93]. The model with the lowest loss is assigned the highest weight, reflecting its reliability in making accurate predictions. The final prediction is then obtained through a weighted sum of the posterior predictive distributions for the individual models. This combination of predictions is carefully weighted to minimize the overall divergence from the true distribution, resulting in a more robust and accurate mortality risk estimation. By incorporating Bayesian predictive distributions and applying the Stacking technique, we aim to enhance the predictive performance and reduce the impact of model uncertainties, thus improving the utility of mortality prediction models in clinical decision-making.

The remainder of this chapter is organized as follows. Section 6.2 presents methods on stacking predictive distribution. This is followed by application to CIN paediatrics data. Results are presented in Section 6.3 and we conclude with a discussion in Section 6.4.

6.2 Methods

6.2.1 Prognostic factors considered in development of the meta model.

In this chapter, the four models were considered. The four models were identified from previous chapters, and they include RISC-Malawi model and 3 other models developed by Lowlaavar *et al.* 2016. There were seven predictors included in the RISC-Malawi model, three predictors included in model 1,2 and two predictors in model 3 of the Lowlaavar *et al.* as shown in Table 6.2.1-1. Based on these predictors we developed 4 new models for the model averaging experimentation.

Table 6.2.1-1: Prognostic factors included in the 4 predictive models.

Study and the model's name from which prognostic factors were obtained	Model name	Prognostic factors
<i>Hooli et al. 2016</i>	Model 1	Moderate hypoxemia, severe hypoxemia, moderately malnourished, severely malnourished, child-sex(female), wheezing, and unconsciousness
<i>Lowlaavar et al. 2016</i>	Model 2	Abnormal Blantyre Coma Score, Positive HIV, Weight for age z-score
	Model 3	Abnormal Blantyre Coma Score, Positive HIV, Middle Upper Arm Circumference (MUAC)
	Model 4	Abnormal Blantyre Coma Score, MUAC

6.2.2 Model computations

We used Bayesian analysis in the computations because it is easy to express model uncertainties, and ease of generating densities such as a posterior predictive distribution. We used Stan, a probabilistic programming language written in C++, for obtaining full Bayesian inference through its interface *rstan* in R statistical programming language [143]. Stan uses a No-U-Turn sampler which is a Monte Carlo Markov Chain (MCMC) algorithm to draw samples from a desired distribution by building a Markov-chain of accepted values (out of proposed values) for the unknown parameter as a posteriori distribution, a Monte Carlo Markov Chain (MCMC) algorithm that allows for quicker convergence to a target distribution compared to the Gibbs sampler, which uses a random walk algorithm.

6.2.3 Model fitting

We split the data into train and test set. The training set was used to fit logistic regression models predicting in-hospital paediatric mortality. The test set was used to assess model performance and to generate posterior predictive distribution. For all models we specified four chains which is considered adequate, each with 2000 iterations half of which were devoted to the warm-up (adjusting the sampler's behaviour) and were automatically discarded before results were displayed. To assess model convergence, we performed Gelman-Rubin diagnostics, including visual inspection of the model chains of estimated parameters [143, 144].

6.2.4 Model averaging methods

The classical stacking technique also called stacking of means entails combining models by minimizing the mean squared error of the point estimate, or forming a meta model by combining multiple base learners e.g. Random Forest (RF), Support Vector Classification (SVC), K-Nearest Neighbors (KNN), and Light Gradient Boosting Machine (LGBM), Bootstrap aggregating (Bagging), and Adaptive Boosting (AdaBoost) [142]. In this chapter,

we considered ensemble averaging techniques which entails averaging models' point predictions and averaging model's predictive distributions [93]. Each of these model averaging techniques allows incorporation of several competing models $M = m_1, \dots, m_k$ in the estimation process as described below.

1. *Averaging point predictions*

We used the test set to obtain the patient's predicted risk of in-hospital mortality for each of the models, which was computed using the following logistic function.

$$P(Y = 1|X) = \frac{\exp(\beta_0 + \beta_1 X_1 + \dots + \beta_k X_k)}{1 + \exp(\beta_0 + \beta_1 X_1 + \dots + \beta_k X_k)},$$

where β_0 is the model intercept and $\beta_{1,\dots,k}$ are the regression coefficients for prognostic factors $X_{1,\dots,k}$. We then averaged predictions from all four models for each patient to obtain a weighted in-hospital mortality risk prediction.

2. *Stacking predictive distributions*

We started by first fitting each model to the data using Bayesian inference. After which we calculated model's posterior predictive distributions which is the distribution of predictions that the model would make for new data. To combine the predictive densities, we used a loss function called Kullback–Leibler divergence to determine model weights such that the model with lowest loss is given the most weight [93]. The procedure of stacking model is described as follows; let S be the scoring rule to determine model weights (ω_k) which are obtained by finding a solution to the optimization problem of minimizing the difference between expected predictive density of the aggregated predictive distribution and the true data generative distribution as shown in equation (6.2.4-1).

$$\arg \max_{\omega} S \left\{ \sum_{i=1}^k \omega_k p(\tilde{y}|y, M_k), p_t(\tilde{y}) \right\}, \text{ s. t. } \sum_{k=1}^M \omega_k = 1, \omega_k \geq 0, \quad (6.2.4-1)$$

where $p(\tilde{y}|y, M_k)$ is the predictive distribution of the out-of-sample data \tilde{y} for the model M_k trained on data y , and $p_t(\tilde{y})$ is the unknown true data distribution. The weights (ω_k) are a vector that sum to 1, so they form a simplex. This means that the weights represent a probability distribution over the models.

Therefore, the estimated aggregated predictive density takes the form of $\hat{p}(\tilde{y}|y) = \sum_{i=1}^k \hat{\omega}_k p(\tilde{y}|y, M_k)$. Due to computational challenge the predictive density $p(\tilde{y}|y, M_k)$ is approximated as $p(y_i|y_{i-1}, M_k)$ by the Leave-One-Out cross-validation (LOO) via Pareto-smoothed importance sampling (PSIS) [145, 146]. The calculated weights were then used to combine model's predictive distributions as shown in equation (6.2.4-2).

$$\arg \max_{\omega} \frac{1}{n} \left\{ \sum_{i=1}^n \log \sum_{k=1}^M \omega_k p(y_i|y_{i-1}, M_k) \right\}, \text{ s. t. } \sum_{k=1}^M \omega_k = 1, \omega_k \geq 0 \quad (6.2.4-2)$$

6.2.5 Model performance assessment

Two metrics were used to assess model performance: discriminatory ability and model calibration. The discriminatory ability was determined using the c-statistic (value 0-1, discriminative if > 0.7) [84, 85], while the calibration was measured using the calibration slope and calibration intercept. Slope values near 1 indicated better accuracy while values < 1 indicated extreme risk predictions [88].

6.3 Results

6.3.1 Characteristics of the cohort used to train and test models.

In this analysis, we examined a substantial cohort of 19,117 patients from the CIN database. Out of these, 12,745 patients were designated to the training set, while the remaining 6,372 patients were utilized for testing the performance of our predictive models. The overall mortality rate in the entire cohort was 6.3%, resulting in a total of 1,198 recorded deaths. We also identified a subset of patients (5.6%) who were in an unconscious state. Among these unconscious patients, 25.1% died. It is noteworthy that the mortality rates among unconscious patients exhibited considerable variability across different hospitals, ranging from 14.7% to 35%. Abnormal oxygen saturation levels were prevalent among a significant portion of the patients, with 10.3% (1,970 individuals) exhibiting severe hypoxemia. Alarmingly, the case fatality rate among these patients stood at 17%, emphasizing the critical nature of this indicator in predicting mortality risk. Finally, a small fraction of patients (1.3%) was identified as being HIV+ in the cohort, and this proportion was consistent across both the test and train sets shown in Table 6.3.1-1.

Table 6.3.1-1: cohort of patients used to train and test models.

Indicator	Test (N=6372)	Train (N=12745)	Overall (N=19117)
Mortality	396(6.2%)	802(6.3%)	1198(6.3%)
Child-sex (Female)	2612(41.0%)	5290(41.5%)	7901(41.3%)
Unconscious	348(5.5%)	720(5.6%)	1068(5.6%)
Wheezing	223(3.5%)	450(3.5%)	673(3.5%)
Severely malnourished	326(5.1%)	669(5.2%)	995(5.2%)
Moderately malnourished	1469(23.1%)	2950(23.1%)	4419(23.1%)
Severe Oxygen saturation	650(10.2%)	1320(10.4%)	1970(10.3%)
Moderate oxygen saturation	539(8.5%)	1129(8.9%)	1668(8.7%)
Abnormal Blantyre coma score	640(10.0%)	1317(10.3%)	1957(10.2%)
HIV+ diagnosis	84(1.3%)	173(1.4%)	257(1.3%)

6.3.2 Result of model's stacking weights

The weights reflect the relative importance of each model's predictions in the final ensemble. Notably, the procedure of determining model weights assigned a weight of 0.000 to Model 4 suggesting that its predictions were excluded entirely from the final stacked densities. This decision might have been made based on the model's performance or redundancy with other models, rendering it less valuable in contributing to the final prediction. On the other hand, Model 1 received the highest weight of 0.711, signifying that its predictions held the most significant influence on the ensemble's final prediction. Models 2 and 3 were also included in the ensemble, contributing with weights of 0.214 and 0.075, respectively, although to a lesser extent than Model 1.

The distribution of predicted probabilities for different models is provided in Figure 6.3.2-1. Model 1 has the most concentrated distribution, with most of the predictions falling between 0.5 and 0.75. Model 2 has a wider distribution, with predictions ranging from 0 to 1. Model 3 has a similar distribution to Model 2, but with more predictions at the lower end of the

spectrum. Model 4 has the most spread-out distribution, with predictions ranging from 0 to 1 and a peak at around 0.25. The ensemble model has a distribution that is similar to Model 1, but with slightly more predictions at the lower end of the spectrum.

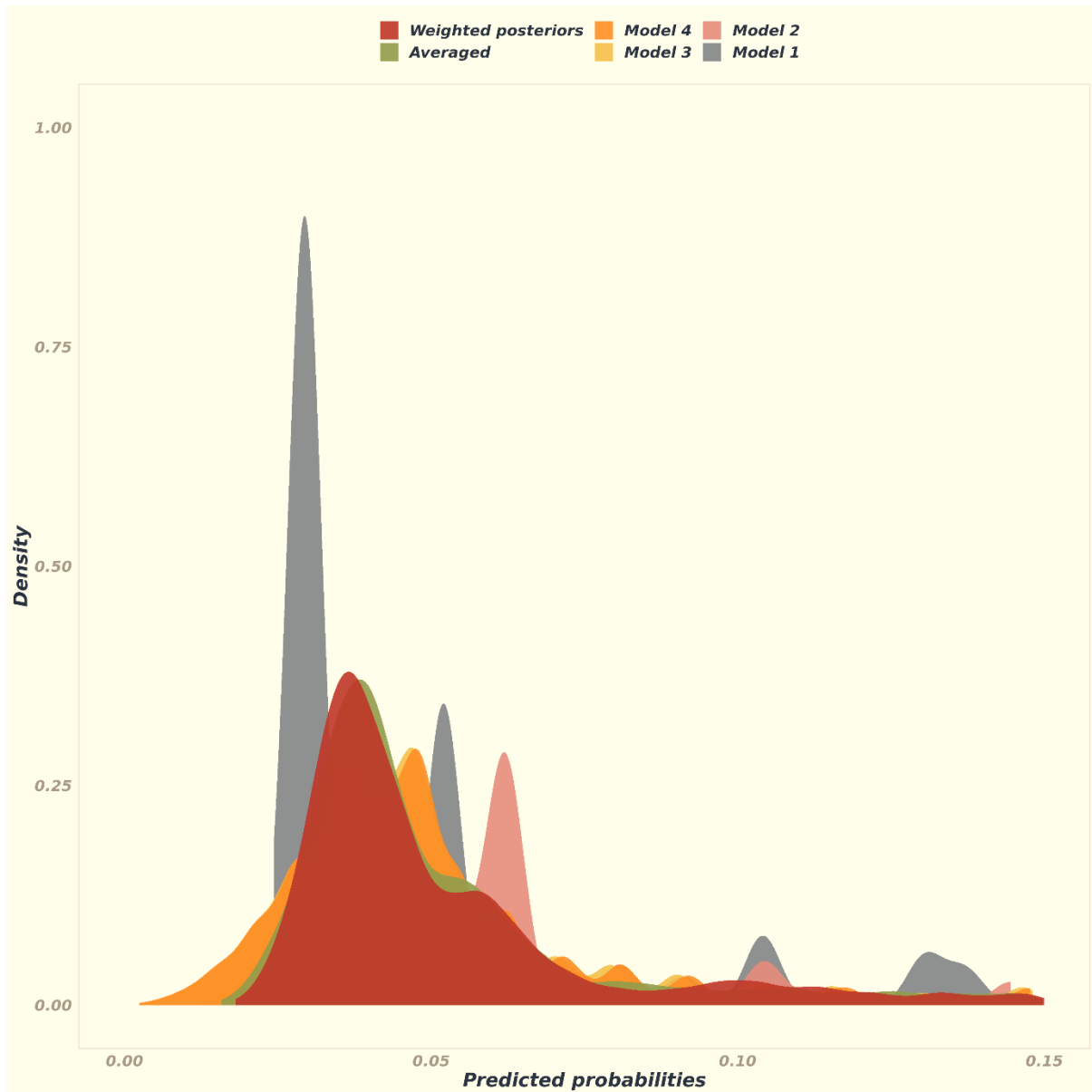


Figure 6.3.2-1: distribution of predicted probabilities for four different models. The x-axis represents the predicted probability, and the y-axis represents the density of predictions at each probability.

6.3.3 Comparison of model performances

i) Discrimination

The discrimination ability (C-statistic) measures the models' capability to accurately distinguish between patients who experienced mortality and those who did not. Higher C-statistic values indicate better discrimination performance, with values closer to 1.0 reflecting excellent predictive accuracy. As shown in Table 6.3.3-1, among the individual models, Model 1 demonstrated the highest discrimination ability with a C-statistic of 0.722, followed by Models 3, 2, and 4, with C-statistics of 0.711, 0.697, and 0.706, respectively. Both ensemble techniques showed improved discrimination abilities compared to the individual models. The Stacking of Predictive Distributions achieved the highest discrimination ability among all approaches, with a C-statistic of 0.744. The Averaging Approach also exhibited a notable discrimination ability, with a C-statistic of 0.740.

Therefore, in summary, the ensemble techniques (Stacking of Predictive Distributions and Averaging Approach) outperformed the individual models in terms of discrimination ability, indicating that they are more accurate in classifying patients based on their mortality risk.

ii) Calibration

The calibration intercept of the individual models ranged from 0.009 to -0.014. A calibration intercept of 0 indicates perfect calibration, while positive values suggest overestimation and negative values suggest underestimation. As shown in Table 6.3.3-1, Model 1 had a calibration intercept closest to ideal calibration (0.009), indicating a slight overestimation of predicted probabilities. The calibration intercepts of the ensemble techniques were -0.005 for Stacking and -0.006 for Averaging. Both ensemble techniques achieved calibration intercepts closer to 0 compared to the individual models, indicating better calibration. This suggests that the ensemble techniques' predicted probabilities were more accurate and closely aligned

with the true probabilities of mortality events. The calibration slopes of the individual models ranged from 0.954 to 1.065. A calibration slope of 1 indicates perfect calibration, where predicted probabilities match the observed probabilities. Values greater than 1 indicate that predicted probabilities are too extreme, while values less than 1 indicate conservative predictions. The calibration slopes of the ensemble techniques were 1.297 for Stacking and 1.266 for Averaging. Both ensemble techniques had calibration slopes greater than 1, indicating a slightly more aggressive prediction of probabilities. However, the calibration slopes were still relatively close to 1, suggesting that the ensemble techniques' predictions were well-calibrated and appropriately reflected the actual probabilities of mortality events.

Table 6.3.3-1: Individual model performances compared to that of ensemble methods.

Model	Calibration intercept	Calibration slope	Discrimination (c-statistics)
Model 1	0.009(-0.097 – 0.115)	0.954(0.848 – 1.060)	0.722(0.682 – 0.758)
Model 2	-0.007(-0.111 – 0.097)	1.060(0.928 – 1.192)	0.697(0.662 – 0.730)
Model 3	-0.013(-0.1108 – 0.092)	1.065(0.942 – 1.188)	0.711(0.681 – 0.740)
Model 4	-0.014(-0.119 – 0.090)	1.041(0.917 – 1.165)	0.706(0.675 – 0.734)
<i>Averaging individual predictions</i>	-0.006(-0.110 – 0.098)	1.266(1.134 – 1.398)	0.740(0.712 – 0.767)
Stacking of Predictive Distributions	-0.005(-0.109 – 0.099)	1.297(1.163 – 1.430)	0.744(0.716 – 0.771)

6.4 Discussion

6.4.1 Summary of the findings

In this chapter, we utilized a stacking methodology to merge the predictive distributions originating from four distinct prognostic models. The primary objective was to bolster the precision and dependability of mortality risk predictions concerning pediatric patients, and hence reducing model uncertainty. The procedure of stacking encompassed the allocation of weights to each model, calculated through the application of the Kullback–Leibler divergence method which is a measure used to quantify how one probability distribution differs from another. These weights played a crucial role in delineating the unique contributions of each model towards the ultimate collective prediction. Here is the summary of the key findings: Firstly, the discriminatory ability of the ensemble techniques, namely the Stacking of Predictive Distributions and the Averaging Approach, surpassed that of the individual models. Discrimination ability, as quantified by the C-statistic, reflects a model's capacity to effectively distinguish between patients who experienced mortality and those who did not. While individual Model 1 demonstrated commendable discrimination ability with a C-statistic of 0.722, both ensemble techniques exhibited even better performance. The Stacking of Predictive Distributions achieved the highest discrimination ability, recording a C-statistic of 0.744. Similarly, the Averaging Approach showcased notable discrimination ability, with a C-statistic of 0.740. This robustly indicates that the ensemble techniques excel in accurately classifying patients based on their risk of mortality. Secondly, the ensemble techniques showcased marked improvements in calibration when compared to the individual models. Calibration is pivotal for aligning predicted probabilities with observed probabilities and thus enhancing the accuracy of risk estimation. While the calibration intercept of Model 1 was relatively close to ideal calibration at 0.009, the ensemble techniques further improved upon this metric. Specifically, the calibration intercepts of the Stacking and Averaging approaches

were -0.005 and -0.006, respectively. These values suggest that the ensemble techniques' predicted probabilities were closely aligned with the true probabilities of mortality events. Moreover, the calibration slopes of the ensemble techniques (1.297 for Stacking and 1.266 for ordinary point averaging) indicated a slightly more aggressive prediction of probabilities, while still maintaining proximity to the ideal value of 1. This underscores that the ensemble techniques' predictions were well-calibrated and aptly reflected the actual probabilities of mortality events.

6.4.2 Stacking of predictive distribution method vs other approaches.

Early literature on stacking techniques primarily focused on averaging point predictions, a method sometimes referred to as "stacking of means." It aimed to combine models by minimizing the mean squared error of the point estimate [140, 147]. In earlier research by Raftery *et al.* [137], Clerke *et al.* [140] and Hoeting *et al.* [20], Bayesian Model Averaging was introduced as a potential solution. However, it later proved to be unreliable due to principles such as Occam's Razor and Occam's Window, which implied that it would asymptotically select a single model, the one closest in KL divergence to the true data-generating process. Mortality prediction, a critical outcome in hospitalization contexts, has remained a challenging task. It's important to clarify that mortality prediction involves a convolution of single models, rather than a straightforward mixture, making it impossible for any approach to recover the true model from the list. Here, Bayesian Stacking of Predictive Distributions provides a more advanced ensemble technique. Instead of fixating on point estimates, it considers the complete predictive distributions produced by individual models. Each model's Bayesian predictive distribution encompasses the full spectrum of predictions it could make for new data. The stacking process amalgamates these predictive distributions, using a loss function like the Kullback-Leibler divergence to quantify the differences between these distributions and the actual one. The model with the lowest loss garners the highest

weight, resulting in a final output that is not a single point estimate but a predictive distribution. This approach provides a comprehensive view of model uncertainty, offering insights into a range of potential outcomes and their associated probabilities.

6.4.3 Limitations

This study's limitations are rooted in its use of predictors drawn from established models (RISC-Malawi, Lowlvaal *et al.*), which may constrain predictive capabilities by relying on available variables, potentially overlooking other important prognostic factors. However, it's crucial to clarify that the primary aim of this chapter and the associated research was not to devise a novel predictive model, but rather to introduce an innovative technique that harnesses the posterior predictive distributions of existing models to enhance out-of-sample predictions. Stacking of predictive distribution methods entail a greater computational burden compared to simpler averaging approaches. Furthermore, the effectiveness of stacking is intricately tied to the selection of models, as it cannot outperform the best linear combination offered by the chosen model list[93]. Notably, stacking displays a relatively lower sensitivity to model misspecifications, underscoring its role as a model averaging tool rather than a model selection procedure, with its ultimate focus on enhancing predictive accuracy.

6.4.4 Implications of the study findings

The implications of our study's findings carry substantial significance for both the predictive modeling field and the domain of clinical decision-making, especially in the context of predicting pediatric mortality in hospital settings. Our research illuminates the potential of ensemble techniques to significantly enhance the precision and dependability of mortality risk forecasts for pediatric patients, with profound consequences for clinical practice, healthcare resource allocation, and overall patient care.

In the realm of predictive modeling, our study introduces a pivotal paradigm shift by harnessing the capabilities of ensemble techniques such as the Stacking of Predictive Distributions and the Averaging Approach. These methodologies transcend the constraints of single-model approaches by amalgamating insights from multiple predictive models. This amalgamation results in a more comprehensive and resilient assessment of mortality risk for pediatric patients. Given the critical nature of pediatric mortality, especially in a hospital setting, the heightened discriminative prowess of ensemble techniques becomes paramount. This improved capacity to differentiate between high-risk and lower-risk cases empowers healthcare providers to allocate their attention and resources more efficiently, thereby optimizing patient care strategies.

6.5 Conclusion

The chapter's approach, centered around the stacking of Bayesian predictive distributions from individual models, represents a pivotal advancement in addressing model uncertainty in prognostic modeling. By leveraging the power of Bayesian methods and the insights of ensemble techniques, this methodology offers a robust and reliable framework for predicting in-hospital pediatric mortality. The innovative amalgamation of predictive distributions, coupled with the calculation of model weights through loss functions, results in a final prediction that is more accurate and better calibrated. This improvement in predictive performance holds significant implications for clinical decision-making, enabling clinicians to make informed choices regarding the management and prioritization of pediatric patients who are at an increased risk of deterioration. The approach's potential to reduce the impact of model uncertainties underscores its importance in enhancing the utility of mortality prediction models in clinical practice. In a landscape where every incremental improvement in predictive accuracy matters, this study's contribution marks a significant stride toward more reliable and effective prognostic modeling in pediatric healthcare.

Chapter 7

Quantifying the Impact of Short Follow-Up Period on the prognostic model: A Monte Carlo Simulation Study

7.1 Introduction

Researchers should accommodate for competing risks, which preclude or fundamentally change the probability of the event of interest, to avoid biased estimates of the relationship between covariates and patient outcomes. When the event of interest is in-hospital mortality, competing risks include discharges, referrals, etc. Using the conventional Cox Proportional Hazard(CoxPH) model that censors the competing events may yield biased estimates [148]. It is important to understand how failing to account for competing risks of in-hospital mortality may affect estimation accuracy because some scoring systems employing the standard Cox regression where competing risks are censored have been developed and deployed in practice in other specialities such as nephrology [149].

Fine and Gray developed the proportional Sub-distribution Hazard (SH) model for modelling the effects of covariates on the Cumulative Incidence Function (CIF) in the presence of competing risks [17, 150]. However, in the Chapter 2 of this thesis(systematic review of prognostic models) it was shown that competing risks are ignored partly due to the analytical complexity of techniques involving competing risks [151, 152].

Most paediatric in-hospital mortality occurs soon after admission [99]. However, the application of the SH model in such a setup of a short follow-up period or heavy censoring has been discouraged due to its effect on the proportionality assumption of the SH model [153]. In addition, it has also been argued that the competing risks framework has negligible influence compared to other alternative approaches when the follow-up period is short [18]. Despite this evidence, it is notable that no simulations have been done to quantify the unreliability of the SH model in a setting of heavy censoring or a short follow-up period.

In this chapter, the objective is to examine the accuracy of estimated quantities of the SH model in patients with short follow-up periods through extensive Monte Carlo simulations, which work by iteratively sampling data randomly and incorporating a range of factors to mimic a statistical problem.

7.2 Methods

7.2.1 Monte Carlo simulations

To examine the accuracy of estimated quantities of the SH model in patients with short follow-up periods, we conducted extensive Monte Carlo simulations, which works by iteratively sampling data randomly and incorporating a range of factors to mimic a statistical problem. To do this, we used plasmode-type simulations whereby empirical data analysis informed the design of the simulations [154]. Plasmode is a real dataset that is created from natural processes but has some aspect of the data-generating model that is known. This approach was motivated by the fact that clinical data have complex covariance structures which cannot be replicated in a fully synthetic data [155, 156].

7.2.2 Simulation scenarios

In the simulation, we allowed three factors to vary, namely:

- i) the sample size (n) of the simulated datasets,

- ii) the parameter (p) which is the proportion of patients with covariates equal to zero who experience mortality as time (t) gets arbitrarily large, and
- iii) hospital survival/follow-up period measured by the LOS.

The sample size (n) took three values: 500, 1000, and 10000. The parameter p could take on three values: 0.1, 0.5 and 0.9. LOS could take 15 values ranging from 1 to 15. We thus examined 135 ($3 \times 3 \times 15$) different scenarios and simulated 100 datasets for each scenario.

7.2.3 Data-generating process

As part of the simulation design, we included a set of covariates in the Sub-distribution Hazard (SH) model fitted in the empirical dataset (CIN) and whose model coefficients were used as part of the true data-generating mechanism in the simulation of event time and the type of event. We chose four including age (in months), child sex, the temperature at admission, and presence of severe acute malnutrition (defined as one of the following: mid-upper arm circumference < 11.5 cm, severe wasting, oedema, or clinical admission diagnosis of the severe forms of malnutrition including kwashiorkor, marasmus, or marasmus-kwashiorkor). In the variable selection, we didn't follow any set criteria. However, as part of the simulation design, we wanted to examine the effect of various variable characteristics on the estimation accuracy. Aspects of interest included covariate variability (for continuous) and frequency of observation (for categorical variable). Child sex and body temperature at admission were included because they were frequently observed and less variable, respectively. The opposite was true for severe acute malnutrition and patients' age. Continuous variables were standardized to have a mean zero and unit variance prior to inclusion into the model. As part of the sensitivity analyses in a separate set of simulations, we included unstandardized continuous variables into the model to gain an insight on the effect it would have on the estimation accuracy.

7.2.4 Model coefficients based on the empirical data.

We subdivided the CIN data into separate subpopulations based in the LOS which ranged from 1 to 15. Accordingly, a dataset with LOS=1 included patients admitted and discharged from hospital or died at the same day of admission, while those of LOS=5 included patients with LOS ranging from 1 to 5. Similarly, a dataset of LOS=15 included patients with at most 15 LOS and thus included patients of various LOS ranging from 1 to 15. In each of these 15 datasets, we fitted a sub-distributional hazard model (equation 7.2.4-1) for each of the possible patient outcomes shown in Figure 3.6.1-1

$$\lambda_j(t|\mathbf{z}) = \frac{-\partial \log\{1 - F_j(t|\mathbf{z})\}}{\partial t}, j \in \{1,2,3,4\} \quad (7.2.4-1)$$

where j denotes various possible patient outcomes, \mathbf{z} is the covariate vector, $F_j(t) = \int_0^t P(T > u -) \alpha_{0j}(u) du$, is the cumulative incidence function for the j^{th} event at time t , $P(T > u -)$ is the probability of being in the original state 0 (point of admission) before transiting to state j in time u .

7.2.5 Simulation of event types and time-to-event

For the simulated dataset to be complete, we needed to generate the event-types and the time-to-event. To achieve this, we bootstrapped the empirical data for all four covariates for each scenario. The resultant dataset (scenario replicated dataset) was combined with the model coefficients obtained from the SH model fitted on the empirical dataset to generate a linear predictor ($\mathbf{X}\beta$) for each patient. We simulated time-to-event (T_E) for various types of outcomes using the indirect method by Fine and Gray[17] which has also been described in detail in the Beyersmann book[157]. The primary event of interest (death) was assumed to have the following distribution in equation (7.2.5-2).

$$\Pr(T_i \leq t | Z_T = 1, \mathbf{X}) = \frac{1 - (1 - p(1 - e^{-t}))^{\exp(\mathbf{X}\boldsymbol{\beta})}}{1 - (1 - p)^{\exp(\mathbf{X}\boldsymbol{\beta})}}, \quad (7.2.5-2)$$

where p is the proportion of patients with covariate equal to zero who experience event of interest (death) as time t becomes arbitrarily large, $1 - (1 - p(1 - e^{-t}))^{\exp(\mathbf{X}\boldsymbol{\beta})}$ is the cumulative incidence function of the primary event of interest (death), and $1 - (1 - p)^{\exp(\mathbf{X}\boldsymbol{\beta})}$ is the probability of the occurrence of death as $t \rightarrow \infty$. The inverse of equation 2 yielded equation (7.2.5-1) which was evaluated to generate time to primary event(T_1) as guided by Austine *et al.* [75].

$$T_1 = -\log \left(-\frac{1 - \left(\left(-u + \frac{1}{1 - (1 - p)^{\exp(\mathbf{X}\boldsymbol{\beta})}} \right) (1 - (1 - p)^{\exp(\mathbf{X}\boldsymbol{\beta})}) \right)^{\frac{1}{\exp(\mathbf{X}\boldsymbol{\beta})}} - p}{p} \right), \quad (7.2.5-3)$$

where $u \sim Unif(0,1)$, $\mathbf{X}\boldsymbol{\beta}$ is the linear predictor – a combination of the bootstrapped data and the vector of coefficient of proportional sub-distribution hazard model from equation 1. Time to experience competing event were generated from an exponential distribution $T_j = \exp(\mathbf{X}\boldsymbol{\gamma}_j)$, where, $\mathbf{X}\boldsymbol{\gamma}_j$ is the combination of covariate vector and the coefficient of SH model for various competing events, and j denotes the competing event. In addition, we also generated the observed survival time (T) using this relationship $T = \min(T_j)$ where j denotes patients' outcome. The event type that corresponded with the minimum survival time (T) was assumed to be the appropriate patient outcome.

7.2.6 Searching rate parameter of exponential distribution through bisection method

To find the best possible values of time at which subjects are censored, optimal values of the rate parameters λ_{censor} of the exponential distribution were required. For each scenario and in each iteration a new seed was set that corresponded to the iteration number. Setting of seed ensured the reproducibility (reuse of the same set of random variables) of the sequence where necessary. Exponential random variables were simulated from a huge population (N=1,000,000). This was informed from previous simulation studies. Optimal λ_{censor} values were searched in the interval [0.01, 500] to achieve the desirable proportion of patients for whom an event was observed to occur in the simulated data using a bisection algorithm. This approach starts with a large interval known to contain the solution, then it successively reduces interval size until the solution is found. Our stopping criteria was defined as follows; if the difference between the probability of being censored in the empirical dataset and in the simulated dataset was negligible (<0.0001) for any given scenario. The schematic view of how this was implemented is as shown in Figure 7.2.6-1.

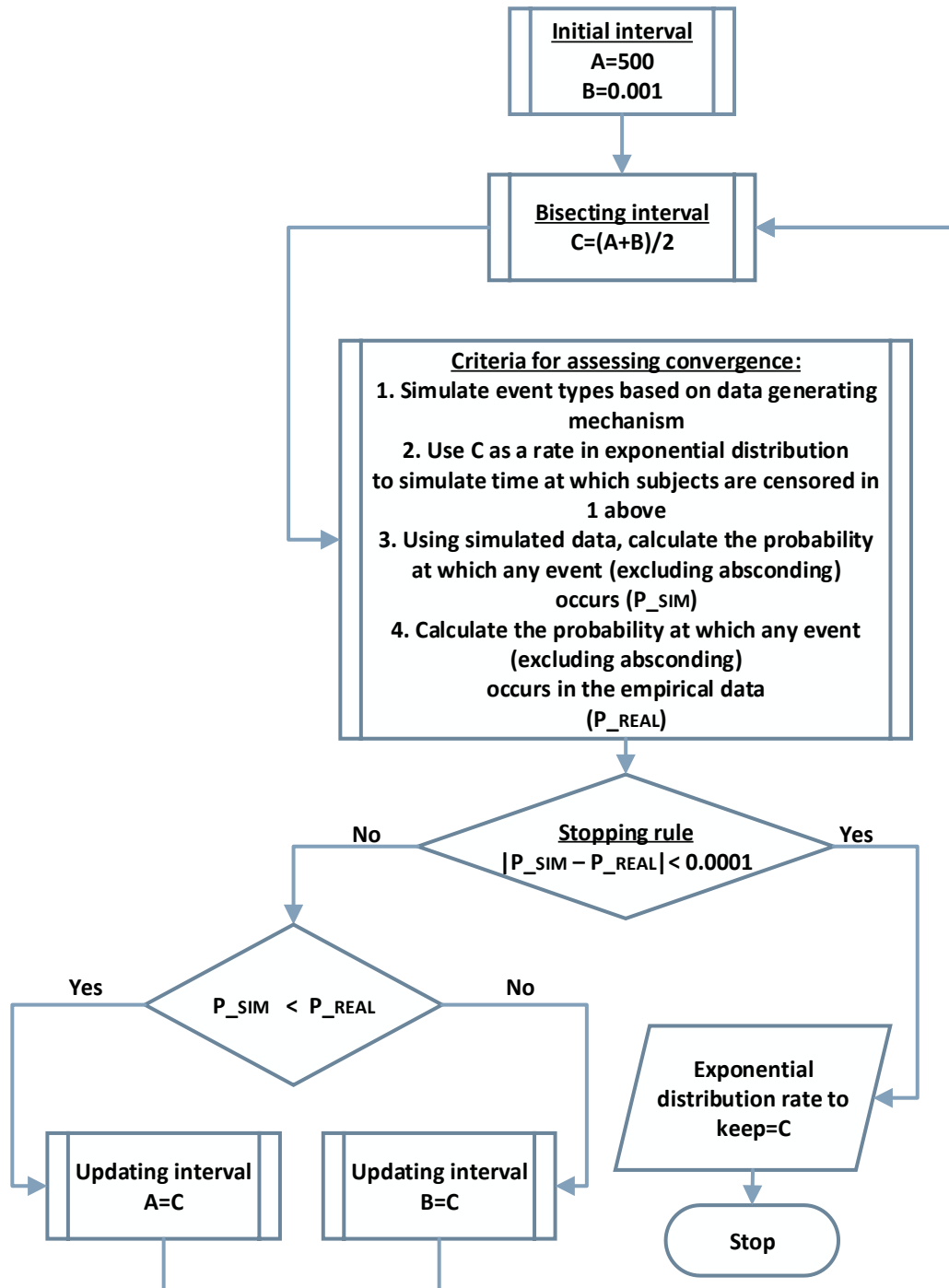


Figure 7.2.6-2: The bisection algorithm used to search for the optimal rate parameter for the exponential distribution that was used to simulate time at which patients were censored.

7.2.7 Model coefficients based on the simulated data.

We added the simulated event type and time-to-event for each patient in the bootstrapped set. We used the resultant dataset to fit the SH model using the primary event (in-hospital death) as the dependent variable and regressed against the four covariates.

7.2.8 Assessing estimation bias

Upon model convergence, we extracted the estimated regression coefficients. We estimated the bias and relative bias for each scenario across the 100 replicated datasets using equation (7.2.8-1)

$$bias(\beta_j) = \frac{1}{100} \sum_{i=1}^{100} (\hat{\beta}_{i,j} - \beta_j), \quad (7.2.8-1)$$

where β_j is the true value of the j^{th} regression coefficient (obtained via the empirical dataset) and $\hat{\beta}_{i,j}$ is the estimated value of the j^{th} regression coefficient obtained in the i^{th} simulated dataset. The relative bias was defined as $\frac{bias(\beta_j)}{\beta_j} \times 100$.

7.3 Results

7.3.1 Characteristics of the empirical sample used for data generation.

In total there were 140,203 patients across the 19 hospitals analysed. The overall mortality was 6.2% which ranged from 1.3% to 11.8% across hospitals as shown in Figure 7.3.1-1. Overall in-patient mortality was 6.2% ranging from 1.3% to 11.8% across hospitals. Subpopulation analysis suggested that mortality was the highest (40.6%) among patients admitted and discharged on the same day (LOS=1), and 60.3% of all 15 days' mortality occurred within the first 48 hours of admission, as shown in Figure 2. The distribution of the four covariates used in the data generating model showed that, 44.1% of children were female, 9.6% had severe acute malnutrition, and the average age in months was 34.4 (standard deviation: 35.7) which included neonates with a median age of 9 days (interquartile range 5-

14 days). Patients' characteristics were similar between patients with short and long follow-up periods. However, severe acute malnutrition was more prevalent in patients who stayed longer in the hospital as shown in Table 7.3.1-1.

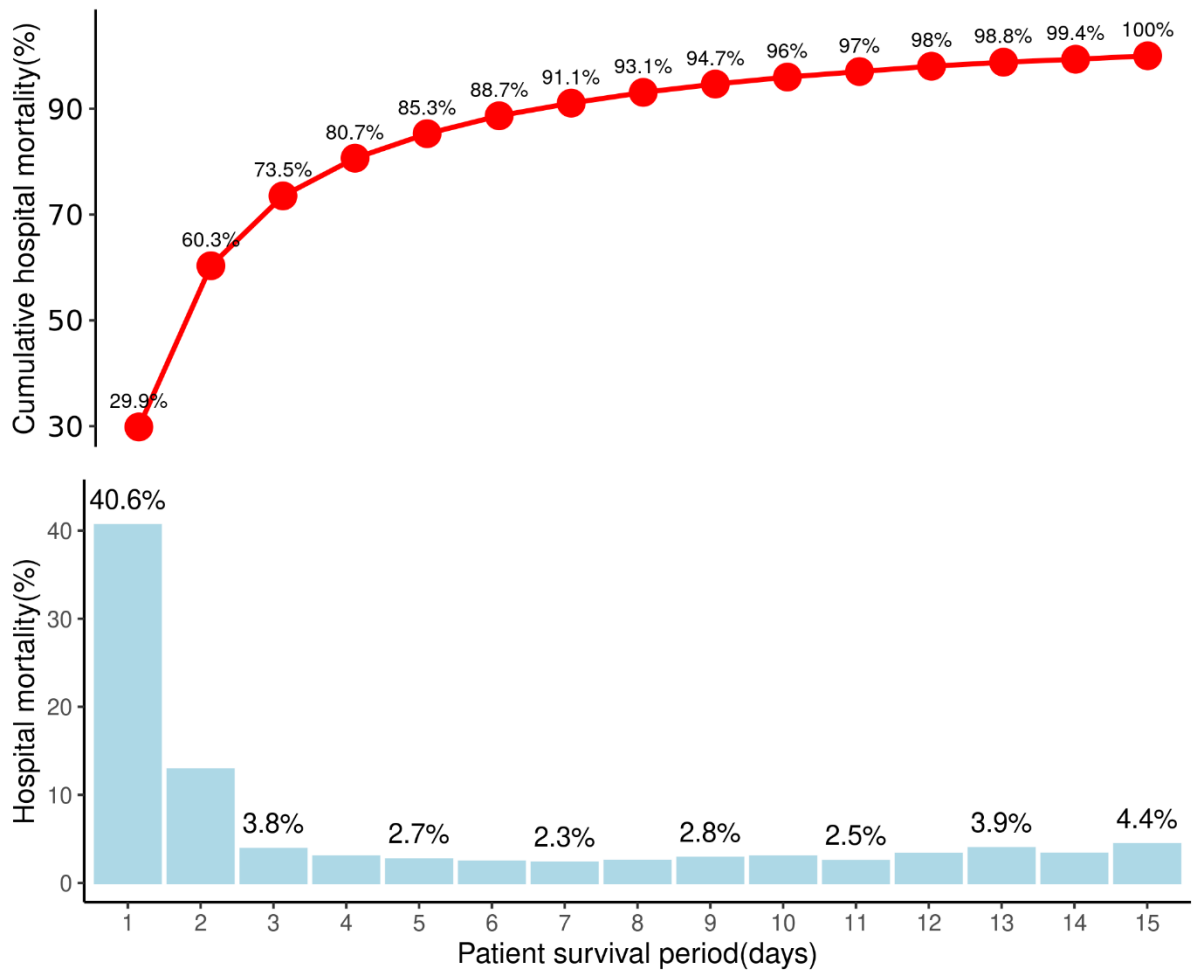


Figure 7.3.1-1: Distributions of the patient outcomes across hospitals

Table 7.3.1-1: Distribution of clinical characteristics of the empirical sample used for data generation.

	Short follow-up (LOS≤2)	Other follow-up period (LOS>2)	All patients
	(N=27018)	(N=113185)	(N=140203)
Outcomes (competing events)			
Discharged/medical stability	20291 (75.1%)	106924 (94.5%)	127215 (90.7%)
Referred out	978 (3.6%)	1875 (1.7%)	2853 (2.0%)
Discharged against medical advice/absconded	338 (1.3%)	560 (0.5%)	898 (0.6%)
Died	5197 (19.2%)	3406 (3.0%)	8603 (6.1%)
Missing	214 (0.8%)	420 (0.4%)	634 (0.5%)
Child sex			
Male	14752 (54.6%)	62311 (55.1%)	77063 (55.0%)
Female	11999 (44.4%)	49890 (44.1%)	61889 (44.1%)
Missing	267 (1.0%)	984 (0.9%)	1251 (0.9%)
Age in months			
Mean (SD)	35.6 (35.1)	34.1 (35.8)	34.4 (35.7)
Missing	230 (0.9%)	960 (0.8%)	1190 (0.8%)
Body temperature at admission (°C)			
Mean (SD)	37.4 (1.26)	37.6 (1.19)	37.6 (1.21)
Missing	3963 (14.7%)	11339 (10.0%)	15302 (10.9%)
Severe Acute Malnutrition			
Yes	1678 (6.2%)	11765 (10.4%)	13443 (9.6%)
No	23236 (86.0%)	96809 (85.5%)	120045 (85.6%)
Missing	2104 (7.8%)	4611 (4.1%)	6715 (4.8%)

7.3.2 Relative bias in the estimated quantities

The results of the Monte Carlo simulations are reported graphically. Model converged in 94.5% of all Monte Carlo simulations whose bias in recovering a true data-generating model are reported graphically. As expected, the estimation accuracy improved with the increase in sample size. The same was true with increasing value of parameter p . This observation was

consistent across all model covariates. We also observed that for binary covariates that were not commonly observed, such as severe acute malnutrition (observed in 9.6% of all patients), the trend did not exhibit any pattern suggestive of reducing estimation bias regardless of the values of parameter p and sample size scenarios used in simulations. See Figure 7.3.2-1.

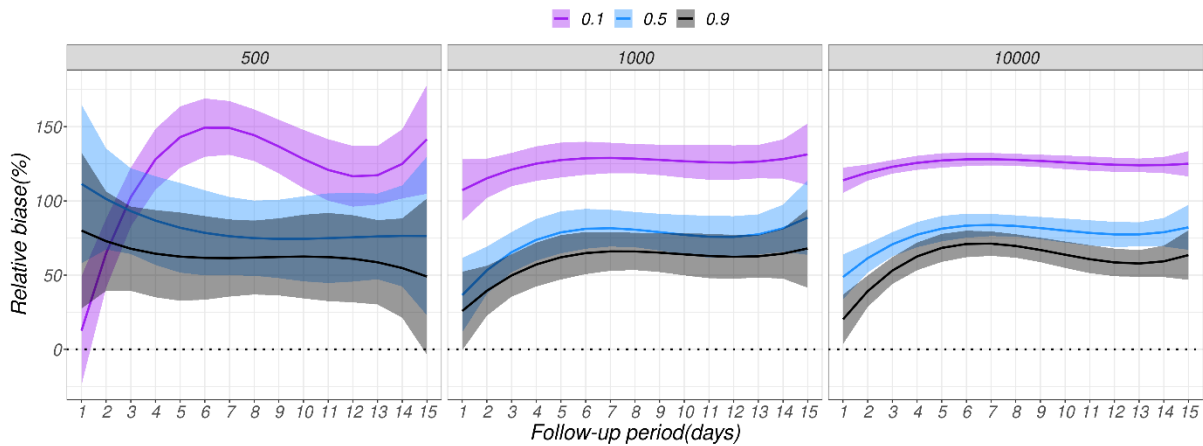


Figure 7.3.2-1: The y-axis is the relative bias in recovering the true model coefficient for variable severe acute malnutrition at different values of the parameter p (0.1, 0.5, and 0.9) across various sample sizes (500, 1000, and 10000).

On the other hand, for the categorical variables that were modestly observed in the empirical dataset, such as gender(male), which was observed in 55% of all patients, the results suggested that the accuracy of the model estimates increased with the increase in the follow-up period as shown in Figure 7.3.2-2

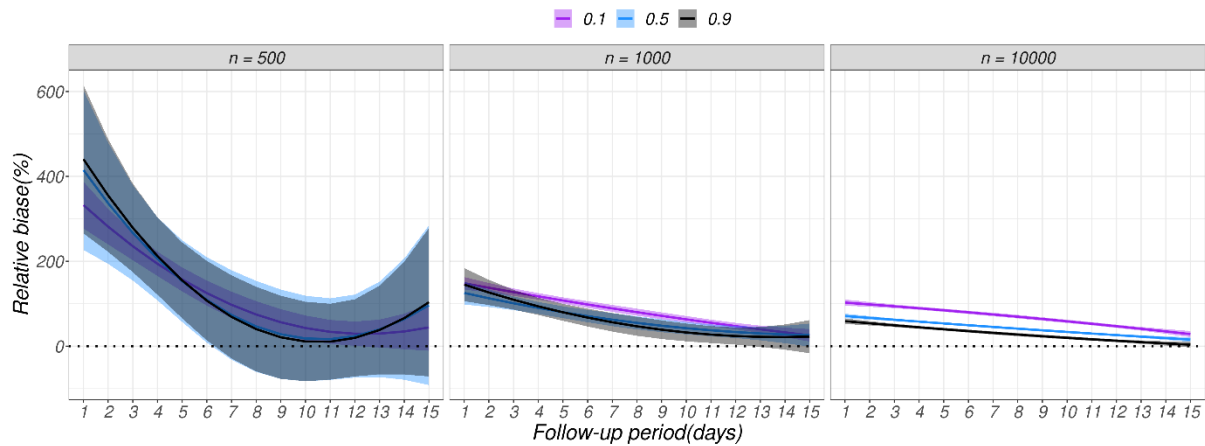


Figure 7.3.2-2: The y-axis is the relative bias in recovering the true model coefficient for variable child sex at different values of the parameter p (0.1, 0.5, and 0.9) across various sample sizes (500, 1000, and 10000).

In the simulations, we included two continuous variables with varied variances: age in months (mean=34.4, standard deviation=35.7) and temperature (mean=37.5, standard deviation=1.2), which were standardized so that they had mean zero and unit variance. However, even after standardizing these variables, we observed that the bias in variable with high variance(age) was relatively high (see Figure 7.3.2-3 lower panel) compared to the temperature variable with a much lower variance as shown in Figure 7.3.2-3 (upper panel). In the sensitivity analyses, we witnessed a high number of model non-convergence (38.2%) owing to non-standardization of continuous variables.

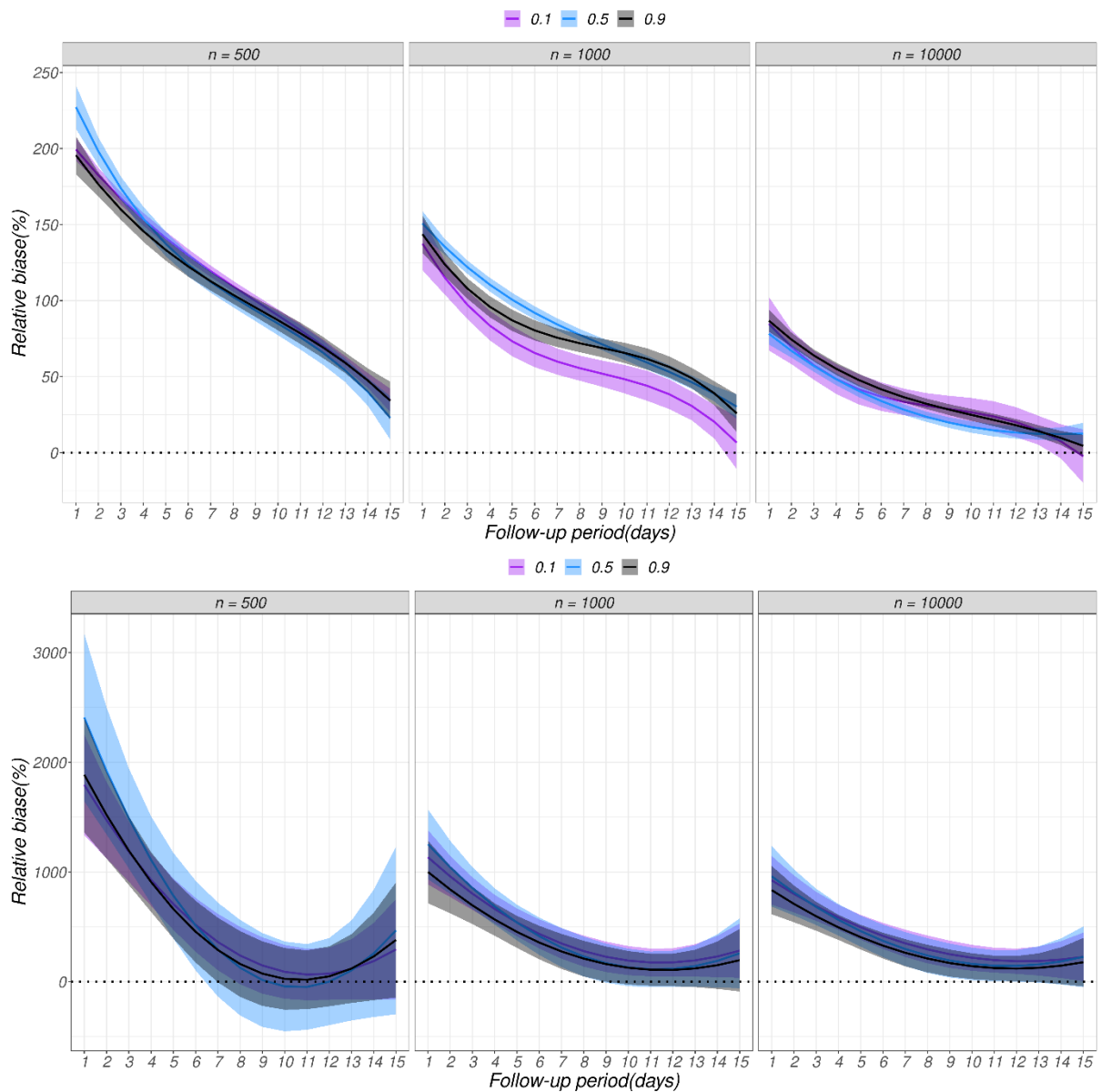


Figure 7.3.2-3: The y-axis is the relative bias in recovering the true model coefficient at different values of the parameter p (0.1, 0.5, and 0.9) across various sample sizes (500, 1000, and 10000). The upper panel represent the variable temperature while the lower panel represent the variable age in months.

7.4 Discussion

7.4.1 Principal findings

The Fine-Gray SH model was developed to model the impact of covariates on the incidence of events over time in the setting of competing risks. In this chapter, through a series of Monte Carlo simulations, we determined how the length of patient follow-up affects the estimation accuracy of the SH model. Most of the patients in our dataset experienced all-cause hospital mortality shortly after admission - 40.6% occurred within 24 hours of admission (*Figure 7.3.1-1*). Simulation findings suggested that the bias in recovering the true data-generating mechanism was relatively higher in scenarios with a short survival period than in long follow-up periods.

Treating competing events as censored observations is commonplace in most epidemiological work, where the CoxPH is used as a standard survival analysis method. While this approach is useful in understanding disease aetiology, it is not reliable if there is a dependence between competing events. This is because the cause-specific hazard cannot be interpreted as the marginal hazard, and covariate effects do not directly translate onto the cumulative scale. One of the difficulties associated with competing events and, in fact, censoring in general is that, without making restrictive assumptions about the exact nature of the dependency between the different event types, it is impossible to distinguish between dependent and independent event processes [158].

While we believe that the bias in recovering the true data-generating mechanism is solely attributable to the follow-up period, we would like to acknowledge that the Breslow method, which is the default method for handling ties in Cox regression and is implemented in the R package used for competing risks, also has some bias as observed by Berger et al. [159], and hence our findings should be interpreted with caution.

The effect of the follow-up period on estimation of regression coefficients for binary covariates with a moderate prevalence was similar to that of continuous covariates. However, for binary covariates with a very low prevalence e.g., 9.6% for severe acute malnutrition, we witnessed a huge variability among model results obtained from simulated data prompting a much higher number of iterations and a longer follow-up period to permit accurate estimates. A similar phenomenon was observed in the sensitivity analyses. This observation is probably because a low-prevalence predictor in the model induces the problem of separation and monotone likelihood in the estimation of parameters [132, 160]. In addition, we noticed high bias in the scenarios with low values of the parameter $p \leq 0.1$. These observations have implications for analysts fitting SH model in which the prevalence of covariate is low or in a setting with low primary event of interest. We witnessed a considerably high number of nonconvergence of models in the sensitivity analyses. This finding underscores the need to standardize continuous covariates before modelling.

7.4.2 Strengths of the study

Methodological issues affecting the SH model, such as the optimal number of events-per-variable (EPV), the impact of various censoring distributions, and the effect of time-dependent covariates, have been addressed previously in multiple studies [75, 161, 162].

To the best of our knowledge, this is the first study to explore how the accuracy of the estimated SH model coefficients is affected by the patients' length of hospital stay.

We used the indirect simulation method developed by Fine and Gray to generate the event types and event times, which was a strength of this work because the method does not require the specification of the cause-specific hazard functions, which should be non-negative. In addition, we used plasmode-type simulations to capture the complex variance-covariance structure inherent in clinical datasets [17, 150].

7.4.3 Limitations of the data

Although we used a large dataset that reflected typical scenarios faced by researchers developing models, we did not evaluate all possible scenarios, including adding highly correlated variables to the model. Another concern was that the simulation results were only based on converged models. Instances of non-convergence were omitted (5.6%). Due to computational challenges, we limited the number of scenarios significantly. For example, we only assessed three scenarios for sample size, three for the values of p , and a limited number of datasets generated per scenario. In a similar simulation study authors generated 1000 datasets per scenario, but in our study, we generated 100 datasets per scenario. Regardless of these limitations, we believe that our research is robust enough and it has pointed areas of concern for further research.

7.4.4 Conclusions

Sub-distribution hazard models have been applied in studies with long follow-up periods, such as cancer studies [163], intensive care studies [164] and nosocomial infections studies [165] but few in studies with short follow-up periods. Monte Carlo simulation results demonstrated how inaccurate SH model estimates are when applied in populations with short survival periods. Based on the study findings, it is challenging to be prescriptive on the average follow-up period a population should have to permit accurate SH model estimates. However, the study has highlighted a potential weakness of the SH model application in the setting of a short survival period that can be a subject of further research.

Chapter 8

Conclusion, Recommendations and Further Research

This report illustrates the practical applications of advanced modeling in developing prognostic models. It follows a structured path, beginning with a comprehensive systematic review in Chapter 2. This review aimed to identify predictive scores for in-hospital mortality among pediatric patients in resource-limited countries. Despite finding twenty-one prognostic models across fifteen studies, the analysis unveiled significant quality concerns. These issues included problems with reporting, handling missing data, univariable analysis for predictor selection, small sample sizes, and inappropriate categorization of continuous predictors. None of the identified models met the criteria for good methodological quality, raising doubts about their predictive capabilities.

In Chapter 4, four eligible prognostic models from Chapter 2 underwent external validation. This involved assessing their discriminatory ability and calibration levels using a diverse population of pediatric patients admitted to 20 hospitals from 2014 to December 2021. While all four models displayed fair discriminatory values (AUC 0.70-0.79), a critical problem emerged. Each of these models consistently underestimated the risk of mortality, as indicated by calibration intercepts greater than zero. This underestimation could lead to the misclassification of high-risk patients. To address these issues, Chapter 5 focused on recalibrating the models for in-hospital mortality prediction. The recalibration aimed to rectify situations where these models either overestimated or underestimated the risk of in-hospital mortality. The use of large sample sizes from 20 county referral hospitals, offering temporal

and spatial richness, was a notable aspect of this work. Two recalibration strategies, calibration-in-the-large adjustment and logistic calibration were explored. Logistic recalibration was found to be more effective, yet the improvements achieved were relatively small. The models still failed to meet the necessary calibration thresholds for clinical use, mainly due to an insufficient account of model uncertainty during their development.

Chapter 6 tackled the issue of model uncertainty. It employed a stacking methodology to merge predictive distributions from four distinct prognostic models (an extension of Chapter 5). This was done with the goal of improving the accuracy and reliability of mortality risk predictions for pediatric patients while reducing model uncertainties. Weights were assigned to each model using the Kullback–Leibler divergence method. Key findings showed that ensemble techniques, including the Stacking of Predictive Distributions and the Averaging Approach, outperformed individual models. They excelled in distinguishing between patients with different levels of mortality risk and improved calibration compared to individual models. This approach's potential to reduce model uncertainties has significant implications for clinical decision-making, resource allocation, and patient care in pediatric healthcare, representing a notable advancement in predictive accuracy and reliability in the field of prognostic modeling.

Chapter 7 shifted the focus to the Fine-Gray Sub-distribution Hazard (SH) model, designed to address competing risks. Through Monte Carlo simulations, it examined how the length of patient follow-up affects the accuracy of this model. The study's findings highlighted that biases in recovering the true data-generating mechanism were more pronounced with shorter follow-up periods. Additionally, it underscored the common practice of treating competing events as censored observations, particularly in epidemiological research using the Cox Proportional Hazard (CoxPH) model. However, this approach becomes unreliable in the presence of dependencies between competing events, as the cause-specific hazard doesn't directly translate to the cumulative scale.

Despite the progress made, there is a need for further research in this area to fully comprehend and address these complexities.

References

1. Hug, L., et al., *National, regional, and global levels and trends in neonatal mortality between 1990 and 2017, with scenario-based projections to 2030: a systematic analysis*. The Lancet Global Health, 2019. **7**(6): p. e710-e720.
2. Rajaratnam, J.K., et al., *Neonatal, postneonatal, childhood, and under-5 mortality for 187 countries, 1970–2010: a systematic analysis of progress towards Millennium Development Goal 4*. The Lancet, 2010. **375**(9730): p. 1988-2008.
3. Ayieko, P., 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*. Archives of disease in childhood, 2015: p. archdischild-2015-309269.
4. Organization, W.H., *Serious childhood problems in countries with limited resources*. 2004.
5. Child, W.H.O.D.o., et al., *Handbook IMCI: integrated management of childhood illness*. 2005: World Health Organization.
6. Moons, K.G., et al., *Prognosis and prognostic research: what, why, and how?* Bmj, 2009. **338**.
7. Maguire, J.L., et al., *Clinical prediction rules for children: a systematic review*. Pediatrics, 2011. **128**(3): p. e666-e677.
8. Bleeker, S., et al., *External validation is necessary in prediction research:: A clinical example*. Journal of clinical epidemiology, 2003. **56**(9): p. 826-832.
9. Van den Bruel, A., et al., *Clinicians' gut feeling about serious infections in children: observational study*. Bmj, 2012. **345**.
10. Selby, D., et al., *Clinician accuracy when estimating survival duration: the role of the patient's performance status and time-based prognostic categories*. Journal of pain and symptom management, 2011. **42**(4): p. 578-588.

11. Wakaba, M., et al., *The public sector nursing workforce in Kenya: a county-level analysis*. Human resources for health, 2014. **12**(1): p. 1-16.
12. Altman, D.G. and P. Royston, *What do we mean by validating a prognostic model?* Statistics in medicine, 2000. **19**(4): p. 453-473.
13. Cox, D.R., *Regression models and life-tables*. Journal of the Royal Statistical Society: Series B (Methodological), 1972. **34**(2): p. 187-202.
14. Kalbfleisch, J.D. and R.L. Prentice, *The statistical analysis of failure time data*. 2011: John Wiley & Sons.
15. Ranganathan, P. and C. Pramesh, *Censoring in survival analysis: potential for bias*. Perspectives in clinical research, 2012. **3**(1): p. 40.
16. Dignam, J.J. and M.N. Kocherginsky, *Choice and interpretation of statistical tests used when competing risks are present*. Journal of Clinical Oncology, 2008. **26**(24): p. 4027.
17. Fine, J.P. and R.J. Gray, *A proportional hazards model for the subdistribution of a competing risk*. Journal of the American statistical association, 1999. **94**(446): p. 496-509.
18. Rothman, K.J., *Epidemiology: an introduction*. 2012: Oxford university press.
19. Attoh-Okine, N.O. and B.M. Ayyub, *Applied research in uncertainty modeling and analysis*. Vol. 20. 2005: Springer.
20. Hoeting, J.A., et al., *Bayesian model averaging: a tutorial (with comments by M. Clyde, David Draper and El George, and a rejoinder by the authors)*. Statistical science, 1999. **14**(4): p. 382-417.
21. Raftery, A.E., D. Madigan, and C.T. Volinsky, *Accounting for model uncertainty in survival analysis improves predictive performance*. Bayesian statistics, 1996. **5**: p. 323-349.
22. World Health Organization, *Serious childhood problems in countries with limited resources*. 2004, Geneva.
23. World Health Organization. *Children: reducing mortality*. 2017 [cited 2018 20 January]; Available from: <http://www.who.int/mediacentre/factsheets/fs178/en/>.
24. World Health Organization, *Handbook IMCI: integrated management of childhood illness*. 2005: Geneva.

25. Vogenberg, F.R., *Predictive and prognostic models: implications for healthcare decision-making in a modern recession*. American health & drug benefits, 2009. **2**(6): p. 218.
26. Bouwmeester, W., et al., *Reporting and methods in clinical prediction research: a systematic review*. PLoS medicine, 2012. **9**(5): p. e1001221.
27. Altman, D.G., *Systematic reviews in health care: Systematic reviews of evaluations of prognostic variables*. BMJ: British Medical Journal, 2001. **323**(7306): p. 224.
28. Ogero, M., R. Sarguta, and S. Akech, *External validation of pediatric prognostic models predicting in-hospital child mortality in resource-limited settings*. BMJ Open, 2022.
29. Moher, D., et al., *Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement*. PLoS medicine, 2009. **6**(7): p. e1000097.
30. WORLD BANK. *World Bank Country and Lending Groups*. 2019 [cited 2019 September]; Available from: <https://datahelpdesk.worldbank.org/knowledgebase/articles/906519-world-bank-country-and-lending-groups>.
31. Shmueli, G., *To explain or to predict?* Statistical science, 2010: p. 289-310.
32. Moons, K.G., et al., *Critical appraisal and data extraction for systematic reviews of prediction modelling studies: the CHARMS checklist*. PLoS medicine, 2014. **11**(10).
33. Wolff, R.F., et al., *PROBAST: a tool to assess the risk of bias and applicability of prediction model studies*. Annals of internal medicine, 2019. **170**(1): p. 51-58.
34. Moons, K.G., et al., *PROBAST: a tool to assess risk of bias and applicability of prediction model studies: explanation and elaboration*. Annals of internal medicine, 2019. **170**(1): p. W1-W33.
35. Reed, C., et al., *Development of the Respiratory Index of Severity in Children (RISC) score among young children with respiratory infections in South Africa*. Plos One, 2012. **7**(1): p. e27793-e27793.
36. Berkley, J.A., et al., *Prognostic indicators of early and late death in children admitted to district hospital in Kenya: cohort study*. BMJ (Clinical Research Ed.), 2003. **326**(7385): p. 361-361.

37. George, E.C., et al., *Predicting mortality in sick African children: the FEAST Paediatric Emergency Triage (PET) Score*. BMC Medicine, 2015. **13**(1): p. 1-12.
38. Helbok, R., et al., *The Lambaréné Organ Dysfunction Score (LODS) is a simple clinical predictor of fatal malaria in African children*. Journal of Infectious Diseases, 2009. **200**(12): p. 1834-1841.
39. Erdman, L.K., et al., *Combinations of host biomarkers predict mortality among Ugandan children with severe malaria: a retrospective case-control study*. Plos One, 2011. **6**(2): p. e17440-e17440.
40. Emukule, G.O., et al., *Predicting mortality among hospitalized children with respiratory illness in Western Kenya, 2009-2012*. Plos One, 2014. **9**(3): p. e92968-e92968.
41. Hooli, S., et al., *Predicting Hospitalised Paediatric Pneumonia Mortality Risk: An External Validation of RISC and mRISC, and Local Tool Development (RISC-Malawi) from Malawi*. PLoS ONE, 2016. **11**(12): p. 1-13.
42. Gallagher, K.E., et al., *The Predictive Performance of a Pneumonia Severity Score in Human Immunodeficiency Virus–negative Children Presenting to Hospital in 7 Low-and Middle-income Countries*. 2019.
43. Dramaix, M., et al., *Prognostic indices for mortality of hospitalized children in central Africa*. American Journal Of Epidemiology, 1996. **143**(12): p. 1235-1243.
44. Olson, D., et al., *Development of a severity of illness scoring system (inpatient triage, assessment and treatment) for resource-constrained hospitals in developing countries*. Tropical Medicine & International Health, 2013. **18**(7): p. 871-878.
45. Rosman, S.L., et al., *Provisional validation of a pediatric early warning score for resource-limited settings*. Pediatrics, 2019. **143**(5): p. e20183657.
46. Collins, G.S.P., 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): p. 55-63.

47. Moons, K.G.M.P., et al., *Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): Explanation and Elaboration*. *Annals of Internal Medicine*, 2015. **162**(1): p. W1-W73.
48. Concato, J., A.R. Feinstein, and T.R. Holford, *The risk of determining risk with multivariable models*. *Annals of internal medicine*, 1993. **118**(3): p. 201-210.
49. Leisman, D.E., et al., *Development and reporting of prediction models: guidance for authors from editors of respiratory, sleep, and critical care journals*. *Critical care medicine*, 2020. **48**(5): p. 623.
50. Vijay Kotu, B.D., *Feature Selection*, in *Data Science Concepts and Practice*, B.D. Vijay Kotu, Editor. 2019. p. 467-490.
51. Debray, T., et al., *A framework for developing, implementing, and evaluating clinical prediction models in an individual participant data meta-analysis*. *Statistics in medicine*, 2013. **32**(18): p. 3158-3180.
52. Kumar, N., et al., *Triage score for severity of illness*. *Indian pediatrics*, 2003. **40**(3): p. 204-210.
53. Concato, J., et al., *Importance of events per independent variable in proportional hazards analysis I. Background, goals, and general strategy*. *Journal of clinical epidemiology*, 1995. **48**(12): p. 1495-1501.
54. Peduzzi, P., et al., *Importance of events per independent variable in proportional hazards regression analysis II. Accuracy and precision of regression estimates*. *Journal of clinical epidemiology*, 1995. **48**(12): p. 1503-1510.
55. Peduzzi, P., et al., *A simulation study of the number of events per variable in logistic regression analysis*. *Journal of clinical epidemiology*, 1996. **49**(12): p. 1373-1379.
56. Fitzgerald, J., R. Azad, and C. Ryan. *A bootstrapping approach to reduce over-fitting in genetic programming*. in *Proceedings of the 15th annual conference companion on Genetic and evolutionary computation*. 2013. ACM.
57. Harrell, F.E., *Regression Modeling Strategies*. BIOS, 2014. **330**.

58. Moons, K.G., et al., *Critical appraisal and data extraction for systematic reviews of prediction modelling studies: the CHARMS checklist*. PLoS medicine, 2014. **11**(10): p. e1001744.
59. Molenberghs, G., et al., *Handbook of missing data methodology*. 2014: CRC Press.
60. Azur, M.J., et al., *Multiple imputation by chained equations: what is it and how does it work?* International journal of methods in psychiatric research, 2011. **20**(1): p. 40-49.
61. Royston, P., D.G. Altman, and W. Sauerbrei, *Dichotomizing continuous predictors in multiple regression: a bad idea*. Statistics in medicine, 2006. **25**(1): p. 127-141.
62. Collins, G.S., et al., *Quantifying the impact of different approaches for handling continuous predictors on the performance of a prognostic model*. Statistics in medicine, 2016. **35**(23): p. 4124-4135.
63. Leisman, D.E., *Rare events in the ICU: An emerging challenge in classification and prediction*. Critical Care Medicine, 2018. **46**(3): p. 418-424.
64. Ensor, J., et al., *Systematic review of prognostic models for recurrent venous thromboembolism (VTE) post-treatment of first unprovoked VTE*. BMJ open, 2016. **6**(5): p. e011190.
65. Smit, H.A., et al., *Childhood asthma prediction models: a systematic review*. The Lancet. Respiratory Medicine, 2015. **3**(12): p. 973-984.
66. Hodgson, L.E., et al., *Systematic review of prognostic prediction models for acute kidney injury (AKI) in general hospital populations*. BMJ open, 2017. **7**(9): p. e016591.
67. Collins, G.S., et al., *External validation of multivariable prediction models: a systematic review of methodological conduct and reporting*. BMC medical research methodology, 2014. **14**(1): p. 40.
68. Collins, G.S., et al., *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): p. 268-277.
69. Fahey, M., et al., *Clinical prediction models for mortality and functional outcome following ischemic stroke: A systematic review and meta-analysis*. PloS one, 2018. **13**(1): p. e0185402.

70. Logullo, P., et al., *Reporting guideline checklists are not quality evaluation forms: they are guidance for writing*. Health Science Reports, 2020. **3**(2).
71. Glasziou, P., et al., *Reducing waste from incomplete or unusable reports of biomedical research*. The Lancet, 2014. **383**(9913): p. 267-276.
72. Kwakkel, G., et al., *Predicting disability in stroke—a critical review of the literature*. Age and ageing, 1996. **25**(6): p. 479-489.
73. Carrillo-Larco, R.M., et al., *Cardiovascular Disease Prognostic Models in Latin America and the Caribbean: A Systematic Review*. Global heart, 2019. **14**(1): p. 81-93.
74. Collins, G.S. and K.G. Moons, *Comparing risk prediction models*. 2012, British Medical Journal Publishing Group.
75. Austin, P.C., A. Allignol, and J.P. Fine, *The number of primary events per variable affects estimation of the subdistribution hazard competing risks model*. Journal of clinical epidemiology, 2017. **83**: p. 75-84.
76. Austin, P.C. and E.W. Steyerberg, *Events per variable (EPV) and the relative performance of different strategies for estimating the out-of-sample validity of logistic regression models*. Statistical methods in medical research, 2017. **26**(2): p. 796-808.
77. Bitwe, R., M. Dramaix, and P. Hennart, *Simplified prognostic model of overall intrahospital mortality of children in central Africa*. Tropical medicine & international health: TM & IH, 2006. **11**(1): p. 73.
78. Ayieko, P., 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): p. 223-9.
79. Ministry of Health (MOH) [Kenya]. *Paediatric Admitting Record Form*. 2015 [cited 2016; Available from: http://www.idoc-africa.org/images/documents/Paeds%20a_%20PAR%20Paediatric%20Admitting%20Record%20Form.pdf.
80. Harris, P.A., et al., *Research electronic data capture (REDCap)--a metadata-driven methodology and workflow process for providing translational research informatics support*. J Biomed Inform, 2009. **42**(2): p. 377-81.

81. Van Buuren, S., *Multiple imputation of discrete and continuous data by fully conditional specification*. *Statistical methods in medical research*, 2007. **16**(3): p. 219-242.
82. White, I.R., P. Royston, and A.M. Wood, *Multiple imputation using chained equations: issues and guidance for practice*. *Statistics in medicine*, 2011. **30**(4): p. 377-399.
83. Moons, K.G., et al., *Using the outcome for imputation of missing predictor values was preferred*. *Journal of clinical epidemiology*, 2006. **59**(10): p. 1092-1101.
84. Steyerberg, E., *Clinical prediction models: A practical approach to development, validation, and updating (statistics for biology and health)* Springer. New York, 2009.
85. Steyerberg, E.W. and Y. Vergouwe, *Towards better clinical prediction models: seven steps for development and an ABCD for validation*. *European heart journal*, 2014. **35**(29): p. 1925-1931.
86. Bijlsma, M.W., et al., *Risk scores for outcome in bacterial meningitis: Systematic review and external validation study*. *Journal of Infection*, 2016. **73**(5): p. 393-401.
87. Muller, M.P., et al., *Can routine laboratory tests discriminate between severe acute respiratory syndrome and other causes of community-acquired pneumonia?* *Clinical infectious diseases*, 2005. **40**(8): p. 1079-1086.
88. Van Calster, B., et al., *Calibration: the Achilles heel of predictive analytics*. *BMC medicine*, 2019. **17**(1): p. 1-7.
89. De Cock, B., et al., *CalibrationCurves: Calibration performance*. 2016.
90. Janssen, K., et al., *Updating methods improved the performance of a clinical prediction model in new patients*. *Journal of clinical epidemiology*, 2008. **61**(1): p. 76-86.
91. Su, T.-L., et al., *A review of statistical updating methods for clinical prediction models*. *Statistical methods in medical research*, 2018. **27**(1): p. 185-197.
92. Steyerberg, E.W., et al., *Validation and updating of predictive logistic regression models: a study on sample size and shrinkage*. *Statistics in medicine*, 2004. **23**(16): p. 2567-2586.
93. Yao, Y., et al., *Using stacking to average Bayesian predictive distributions (with discussion)*. *Bayesian Analysis*, 2018. **13**(3): p. 917-1007.

94. Gray, B., M.B. Gray, and R. Gray, *The emprsk package*. The comprehensive R Archive network, 2004.
95. Emi Suzuki, *Global child mortality rate dropped 49% since 1990*. 2014, World Bank.
96. Shoo, R., *Reducing child mortality: The challenges in Africa*. UN Chronicle, 2007. **44**(4).
97. UNICEF. *Levels and trends in child mortality*. United Nations Inter-Agency Group for Child Mortality Estimation (UN IGME), Report 2021 2021 [cited 2022; Available from: <https://data.unicef.org/resources/levels-and-trends-in-child-mortality/>].
98. Hands, C., et al., *Emergency Triage Assessment and Treatment Plus (ETAT+): adapting training to strengthen quality improvement and task-sharing in emergency paediatric care in Sierra Leone*. Journal of global health, 2021. **11**.
99. Ayieko, P., 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*. Archives of disease in childhood, 2016. **101**(3): p. 223-229.
100. Feinstein, A.R., *Multivariable analysis: an introduction*. 1996, New Haven: Yale University Press.
101. Steyerberg, E.W. and Y. Vergouwe, *Towards better clinical prediction models: seven steps for development and an ABCD for validation*. European heart journal, 2014: p. ehv207.
102. Collins, G.S., et al., *Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement*. BMC medicine, 2015. **13**(1): p. 1.
103. Collins, G.S. and D.G. Altman, *An independent external validation and evaluation of QRISK cardiovascular risk prediction: a prospective open cohort study*. Bmj, 2009. **339**: p. b2584.
104. Sauerbrei, W., et al., *State of the art in selection of variables and functional forms in multivariable analysis—outstanding issues*. Diagnostic and prognostic research, 2020. **4**(1): p. 1-18.
105. Altman, D.G., et al., *Prognosis and prognostic research: validating a prognostic model*. Bmj, 2009. **338**: p. b605.
106. Bleeker, S., et al., *External validation is necessary in prediction research.: A clinical example*. Journal of clinical epidemiology, 2003. **56**(9): p. 826-832.

107. Conroy, A.L., et al., *Prospective validation of pediatric disease severity scores to predict mortality in Ugandan children presenting with malaria and non-malaria febrile illness*. *Critical Care*, 2015. **19**(1): p. 47.
108. Lowlaavar, N., et al., *Pediatric in-hospital death from infectious disease in Uganda: derivation of clinical prediction models*. *PLoS One*, 2016. **11**(3): p. e0150683.
109. Irimu, G., et al., *Tackling health professionals' strikes: an essential part of health system strengthening in Kenya*. *BMJ global health*, 2018. **3**(6): p. e001136.
110. World Health Organization, *Boys weight for age z-score*. 2021.
111. World Health Organization, *Girls weight for age z-score*. 2021.
112. Centers for Disease Control and Prevention, *National Center for Health Statistics, Z.-s.D. Files*, Editor. 2021.
113. Akech, S., et al., *The clinical profile of severe pediatric malaria in an area targeted for routine RTS, S/AS01 malaria vaccination in Western Kenya*. *Clinical Infectious Diseases*, 2020. **71**(2): p. 372-380.
114. Sterne, J.A., et al., *Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls*. *Bmj*, 2009. **338**: p. b2393.
115. White, I.R., P. Royston, and A.M. Wood, *Multiple imputation using chained equations: Issues and guidance for practice*. *Stat Med*, 2011. **30**(4): p. 377-99.
116. Rubin, D.B., *Multiple imputation for nonresponse in surveys*. Vol. 81. 2004: John Wiley & Sons.
117. Florin, T.A., et al., *Reliability of examination findings in suspected community-acquired pneumonia*. *Pediatrics*, 2017. **140**(3).
118. Sjoding, M.W., et al., *Acute respiratory distress syndrome measurement error. Potential effect on clinical study results*. *Annals of the American Thoracic Society*, 2016. **13**(7): p. 1123-1128.
119. Rees, C.A., et al., *External validation of the RISC, RISC-Malawi, and PERCH clinical prediction rules to identify risk of death in children hospitalized with pneumonia*. *Journal of Global Health*, 2021. **11**.

120. Vergouwe, Y., et al., *Substantial effective sample sizes were required for external validation studies of predictive logistic regression models*. Journal of clinical epidemiology, 2005. **58**(5): p. 475-483.
121. Steyerberg, E.W., et al., *Internal validation of predictive models: efficiency of some procedures for logistic regression analysis*. Journal of clinical epidemiology, 2001. **54**(8): p. 774-781.
122. Ogero, M., et al., *Methodological rigor of prognostic models for predicting in-hospital paediatric mortality in low-and middle-income countries: a systematic review protocol*. Wellcome Open Research, 2020. **5**.
123. Ogero, M., et al., *Prognostic models for predicting in-hospital paediatric mortality in resource-limited countries: a systematic review*. BMJ open, 2020. **10**(10): p. e035045.
124. Steyerberg, E.W. and F.E. Harrell, *Prediction models need appropriate internal, internal-external, and external validation*. Journal of clinical epidemiology, 2016. **69**: p. 245-247.
125. Ogero, M., et al., *Examining which clinicians provide admission hospital care in a high mortality setting and their adherence to guidelines: an observational study in 13 hospitals*. Archives of Disease in Childhood, 2020.
126. Finlayson, S.G., et al., *The clinician and dataset shift in artificial intelligence*. The New England journal of medicine, 2021. **385**(3): p. 283.
127. Miller, M.E., S.L. Hui, and W.M. Tierney, *Validation techniques for logistic regression models*. Statistics in medicine, 1991. **10**(8): p. 1213-1226.
128. Phillips, R.S., et al., *Predicting microbiologically defined infection in febrile neutropenic episodes in children: global individual participant data multivariable meta-analysis*. British journal of cancer, 2016. **114**(6): p. 623-630.
129. Nakhjavan-Shahraki, B., et al., *Prediction of clinically important traumatic brain injury in pediatric minor head trauma; proposing pediatric traumatic brain injury (PTBI) prognostic rule*. International journal of pediatrics, 2017. **5**(1): p. 4127-4135.
130. Romanelli, D. and M.W. Farrell, *AVPU score*, in *StatPearls [Internet]*. 2022, StatPearls Publishing.

131. Riley, R.D., et al., *Minimum sample size for developing a multivariable prediction model: PART II-binary and time-to-event outcomes*. *Statistics in medicine*, 2019. **38**(7): p. 1276-1296.
132. Ogundimu, E.O., D.G. Altman, and G.S. Collins, *Adequate sample size for developing prediction models is not simply related to events per variable*. *Journal of clinical epidemiology*, 2016. **76**: p. 175-182.
133. Harrell Jr, F.E., K.L. Lee, and D.B. Mark, *Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors*. *Statistics in medicine*, 1996. **15**(4): p. 361-387.
134. Van Calster, B. and A.J. Vickers, *Calibration of risk prediction models: impact on decision-analytic performance*. *Medical decision making*, 2015. **35**(2): p. 162-169.
135. Pirracchio, R. and O.T. Ranzani, *Recalibrating our prediction models in the ICU: time to move from the abacus to the computer*. 2014, Springer. p. 438-441.
136. Polikar, R. *Ensemble learning*. Scholarpedia 2009; 2776]. Available from: http://www.scholarpedia.org/article/Ensemble_learning.
137. Raftery, A., et al., *Bayesian model averaging in proportional hazard models: Predicting the risk of a stroke*. *Applied Statistics*, 1997. **46**: p. 443-448.
138. Derksen, S. and H.J. Keselman, *Backward, forward and stepwise automated subset selection algorithms: Frequency of obtaining authentic and noise variables*. *British Journal of Mathematical and Statistical Psychology*, 1992. **45**(2): p. 265-282.
139. Surowiecki, J., *The wisdom of crowds*. 2005: Anchor.
140. Clarke, B., *Comparing Bayes model averaging and stacking when model approximation error cannot be ignored*. *Journal of Machine Learning Research*, 2003. **4**(Oct): p. 683-712.
141. Breiman, L., *Stacked regressions*. *Machine learning*, 1996. **24**(1): p. 49-64.
142. Wolpert, D.H., *Stacked generalization*. *Neural networks*, 1992. **5**(2): p. 241-259.
143. Carpenter, B., et al., *Stan: A probabilistic programming language*. *Journal of statistical software*, 2017. **76**(1).

144. Brooks, S.P. and A. Gelman, *General methods for monitoring convergence of iterative simulations*. Journal of computational and graphical statistics, 1998. **7**(4): p. 434-455.
145. Vehtari, A., A. Gelman, and J. Gabry, *Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC*. Statistics and computing, 2017. **27**(5): p. 1413-1432.
146. Vehtari, A., et al., *Pareto smoothed importance sampling*. arXiv preprint arXiv:1507.02646, 2015.
147. El-Rashidy, N., et al., *Intensive care unit mortality prediction: An improved patient-specific stacking ensemble model*. IEEE Access, 2020. **8**: p. 133541-133564.
148. Putter, H., M. Fiocco, and R.B. Geskus, *Tutorial in biostatistics: competing risks and multi-state models*. Statistics in medicine, 2007. **26**(11): p. 2389-2430.
149. Al-Wahsh, H., et al., *Accounting for the competing risk of death to predict kidney failure in adults with stage 4 chronic kidney disease*. JAMA network open, 2021. **4**(5): p. e219225-e219225.
150. Austin, P.C., et al., *Developing points-based risk-scoring systems in the presence of competing risks*. Statistics in medicine, 2016. **35**(22): p. 4056-4072.
151. Beyersmann, J., et al., *Use of multistate models to assess prolongation of intensive care unit stay due to nosocomial infection*. Infection Control & Hospital Epidemiology, 2006. **27**(5): p. 493-499.
152. Beyersmann, J., et al., *Application of multistate models in hospital epidemiology: advances and challenges*. Biometrical Journal, 2011. **53**(2): p. 332-350.
153. Muñoz, A., N. Mongilardi, and W. Checkley, *Multilevel competing risks in the evaluation of nosocomial infections: time to move on from proportional hazards and even from hazards altogether*. Critical Care, 2014. **18**(3): p. 1-4.
154. Franklin, J.M., et al., *Plasmode simulation for the evaluation of pharmacoepidemiologic methods in complex healthcare databases*. Computational statistics & data analysis, 2014. **72**: p. 219-226.
155. Schneeweiss, S., et al., *High-dimensional propensity score adjustment in studies of treatment effects using health care claims data*. Epidemiology (Cambridge, Mass.), 2009. **20**(4): p. 512.

156. Vaughan, L.K., et al., *The use of plasmodes as a supplement to simulations: a simple example evaluating individual admixture estimation methodologies*. Computational statistics & data analysis, 2009. **53**(5): p. 1755-1766.
157. Beyersmann, J., A. Allignol, and M. Schumacher, *Competing risks and multistate models with R*. Use R, ed. K.H. Robert Gentleman, Giovanni Parmigiani. 2011: Springer Science & Business Media.
158. Tsiatis, A., *A nonidentifiability aspect of the problem of competing risks*. Proceedings of the National Academy of Sciences, 1975. **72**(1): p. 20-22.
159. Berger, M., et al., *Subdistribution hazard models for competing risks in discrete time*. Biostatistics, 2020. **21**(3): p. 449-466.
160. Heinze, G. and M. Schemper, *A solution to the problem of monotone likelihood in Cox regression*. Biometrics, 2001. **57**(1): p. 114-119.
161. Donoghoe, M.W. and V. GebSKI, *Impact of the censoring distribution on time-to-event problems in the presence of competing risks*. Trials, 2011. **12**(1): p. 1-1.
162. Poguntke, I., et al., *Simulation shows undesirable results for competing risks analysis with time-dependent covariates for clinical outcomes*. BMC medical research methodology, 2018. **18**(1): p. 1-10.
163. Koller, M.T., et al., *Development and validation of a coronary risk prediction model for older US and European persons in the Cardiovascular Health Study and the Rotterdam Study*. Annals of internal medicine, 2012. **157**(6): p. 389-397.
164. Resche-Rigon, M., E. Azoulay, and S. Chevret, *Evaluating mortality in intensive care units: contribution of competing risks analyses*. Critical Care, 2005. **10**(1): p. 1-6.
165. Wolkewitz, M., et al., *Risk factors for the development of nosocomial pneumonia and mortality on intensive care units: application of competing risks models*. Critical Care, 2008. **12**(2): p. 1-9.