2017

Developing pediatric prognostic model using finite mixture models

Morris Ondieki Ogero
Strathmore Institute of mathematical Sciences (SIMs)
Strathmore University

Developing Pediatric Prognostic Model Using Finite Mixture Models

---

Ogero Morris Ondieki

090034

Submitted in partial fulfilment of the requirements for the Degree of Master of Science in Statistical Sciences at Strathmore University.

Institute of Mathematics (IMS)

Strathmore University

Nairobi, Kenya.

June, 2017

Declaration

I declare that this work has not been previously submitted and approved for the award of a degree by this or any other University. To the best of my knowledge and belief, the thesis contains no material previously published or written by another person except where due reference is made in the thesis itself.

Ogero Morris Ondieki

Signature: _____

Date: _____

Approval

The thesis of **Ogero Morris Ondieki** was reviewed and approved by the following:

Prof. Thomas Achia,

Lecturer, Institute of Mathematical Sciences,

Strathmore University

Ferdinand Othieno,

Dean, Strathmore Institute of Mathematical Sciences,

Strathmore University

Prof. Ruth Kiraka,

Dean, School of Graduate Studies,

Strathmore University

Abstract

Background: World Health Organization (WHO) guidelines recommend early identification of patients who have emergency features for early medical intervention with the aim of reducing child mortality and morbidity. Prognostic models have been developed to be used in clinical setups, but their performance in external validations has been dismal. These poor performances have been attributed to suboptimal statistical methods used for derivation of these scores.

Methods: The Bayesian finite mixture model was used to succinctly identify subpopulations in a population of 47,596 patients from different geographical regions. Mixed models were used to derive a final prognostic model taking into account subgroups of the population. Clinically relevant yet routinely available prognostic factors were used in model development.

Results: Amongst the 23 risk factors used, the AVPU scale which measures unconsciousness was the strongest predictor of mortality with odds of (AOR=2.94, 95% CI= 2.57 - 3.36). Oedema (AOR= 2.66, 95% CI= 2.18 - 3.24), pallor (AOR=2.09, 95% CI= 1.86 - 2.36) and the presence of $>= 3$ severe comorbidities (AOR=2.19, 95% CI= 1.73 - 2.74) were also associated with an increased risk of death.

Conclusion: Given that patient are not alike, a statistical methodology that clusters patients into homogeneous subpopulations should be used to account for the inherent variability in the medical patients. Computational methodology such as mixture models should be used to identify inherent subpopulations that underlie the population of medical patients under study.

Limitation: The use of diagnostic episodes as one of predictors in the model was based on the clinician's impression (not a laboratory test) thus the possibility of false positives could not be ruled out.

Contents

List of Figures

List of Tables

Acknowledgements

I am grateful to my thesis supervisor Prof. Thomas Achia for his invaluable and continuous guidance from the conception throughout the stages of this thesis.

I sincerely appreciate Prof. Mike English and Dr. Philip Ayieko for their steadfast support and insightful ideas in the analysis of the clinical data.

I am grateful to the Clinical Information Network (CIN) team, the participating hospitals and other stakeholders for letting us use their data with minimal restrictions.

I must also acknowledge the support and advice from my colleagues both at KEMRI Wellcome Trust Research Programme (KWTRP) and at Strathmore University for their unflagging support and advice in putting together ideas around this thesis.

Dedication

This thesis is dedicated to:

The memory and friendship of my father, he mentored me throughout life and inspired me to greater heights. He lived his life well, acting upon his spiritual beliefs conscientiously by assisting both friends and strangers in need. He will forever be in my heart.

My mum, brother John Okeri and other family members and also to my friends for being patient with me and supporting me as I worked on this thesis.

List of Abbreviations

Abbreviation	Full name
LIC	Low Income Countries
ML	Machine Learning
AVPU	Alert Verbal Pain Unresponsive
IMCI	Integrated Management of Childhood Illness
PEWS	Pediatric Early Warning System score
GMM	Gaussian Mixture Model
FMM	Finite Mixture Model
MCMC	Markov Chain Monte Carlo
NUT	No-U-TURN
ADVI	Automatic Differential Variational Inference
HMC	Hamiltonian Monte Carlo
WAIC	Watanabe-Akaike Information Criterion
LOO	Leave-One-Out
REDCap	Research Electronic Data Capture
KWTRP	Kenya Medical Research Institute - Wellcome Trust Research Programme

Chapter 1

Introduction

1.1 Background of the study

Childhood mortality has dropped by a good margin from 11 million in 1990 to 6 million in 2013 since the implementation of Millennium Development Goal 4 as shown in (Rajaratnam et al., 2010). Despite the success, mortality still remains a challenge in Low Income Countries (LIC henceforth) with statistics showing that nearly half of under-five deaths in 2012 occurred in sub-Saharan Africa as observed by (Alkema et al., 2016). This challenge seems persistent despite having in place other noble efforts of reducing mortality such as Integrated Management of Childhood Illness (IMCI) (WHO, 2005).

Paediatric scoring systems have been developed in well-resourced countries and have been used to describe the severity of illness in paediatric wards. Examples include the Pediatric Early Warning System score (PEWS) derived by (Duncan et al., 2006) which has helped not only in identifying deteriorating patients in time, but also it has been used in assessing case-mix differences in different clinical trials. Ideally, prognostic or even diagnostic scores are not meant to substitute a clinician's decision in the health facilities, but rather they are meant to augment their judgment and hence used as job-aids in an emergency setup. This is particularly important considering that patient-to-clinicians ratio in LIC has been shown to be high in the paper by (Wakaba et al., 2014).

In LIC, recent studies such as (Ayieko et al., 2015) have shown that hospital mortality often occurs within the first few hours of admission. Therefore, predicting pediatrics' outcome is increasingly becoming important not only in prognosis, but also in planning for

occupancy, staffing and assessment of interventions in the health system research. This has led to an avalanche of studies aimed at developing clinical prognostic models in order to rapidly identify patients who are at a higher risk for the purposes of prioritization in administering treatment in time.

Prediction of paediatrics's outcome has been of interest as evidenced by a vast majority of published scores. However, most of the these published scores such as (Pollack et al., 1988; Shann et al., 1997; Egdell et al., 2008; George et al., 2015; Helbok et al., 2009) are highly specific to a particular pathogen. Common examples of such pathogens include meningococcal disease and malaria which requires a confirmatory laboratory test. This requirement of laboratory screening makes their utility limited especially in an emergency setup. Furthermore, laboratory equipments in most of LIC countries are limited as shown by (Ayieko et al., 2015). Other published scoring systems are faced with numerous challenges such as paltry sample size and a limited events-per-variable which renders such models to overfit and hence predict poorly as demonstrated by (Ogundimu et al., 2016; Steyerberg and Vergouwe, 2014).

Typically, the usefulness of a clinical predictive model is measured by its ability to make correct predictions about future or yet unseen observations. As a result, before any developed model is put into use, it is a conventional requirement as shown by (Bleeker et al., 2003) that an independent external validation to be undertaken. The external validation is often treated preferentially to internal validation of a prediction model. This is because it addresses generalizability of a derived model rather than reproducibility of the same model as asserted by (Steyerberg and Vergouwe, 2014). Essentially, a fully external validation study entails independent researcher(s) studying patients of different geographical location from the one used during model derivation. Performance of the model is assessed by how well it discriminates patients with outcome from those without. A good number of developed paediatric scores have not been generalized because of their poor performance in external validations. Of note, however, almost all developed prognostic models have blatantly failed to account for the uncertainty in model selection as recommended by (Hoeting et al., 1999). Furthermore, they have failed to recognize that there exists sub-populations with different mortality patterns. Inability to count for such uncertainty leads to a model performing dismally during external validation and thus affecting model transportability.

There is therefore need for a practical prognostic model which is based on hospital routine measures collected at the bedside that has intuitive range and that does not need a specific disease or infection to be identified. Such a model would be useful in the real-time discrimination of patients with high risk of deterioration from those with low risk while taking subpopulations into account.

1.2 Problem definition

Clinical prognostic models often combine multiple prognostic factors to give an insight into the relative effects of risk factors in the outcome of interest. Although publications that present such models are becoming more frequent in the literature, the methodology employed is often suboptimal. Consequently, poor performance of the derived models is often observed during their external validations. A model becomes useful if it is able to be generalized and used in the wider population from which the sample is drawn.

In assessing the association between risk factors and patient outcomes, a logistic regression methodology is often used as a standard method. Under this methodology, a *stepwise regression* technique is mostly used to choose a subset of relevant variables despite being discouraged by (Wynants et al., 2016) and (Ogundimu et al., 2016). Inferences about the predictors are then made based on the chosen model constructed of only those variables retained in that model with the assumption of being a unimodal population. However, in practice multi-modal populations are more common to find than unimodal ones, particularly in complex phenomena such as pediatric mortality. Therefore, stepwise method subsequently ignores subpopulations that are inherent in the populations and whose true data-generating mechanism might be very different from one another.

This limitation may be addressed by adopting a finite mixture model approach, which is able to utilize researcher's prior knowledge to come up with finite subpopulations whose statistical properties are unique to each other. This approach has been shown to account for uncertainty in the derived model.

Berkley et al. (2003) observed that paediatrics often experience mortality in different ways; either immediate mortality (on admission), early mortality (few hours after admission $<$ 48 hours) or late mortality ($\geq$ 48 hours after admission). However, a true hospital length of stay (LOS) in terms of hours is largely unknown. Hence a true data generating distribution of pediatric mortality becomes less intuitive because of its multi-modal nature. These subpopulations has to be identified and accounted for so as to make a predictive model useful in out-of sample data.

This study, therefore applied mixture model theory, both as a proof of concept and as an operational model to identify finite subpopulations so as to build an all encompassing

prognostic model. The outcome of interest for this study is all-cause in-hospital mortality. The principle motivation of using mixture models is that the event of interest(mortality) being observed cannot be fully modeled by a single, simple model nor characterized simply by distribution, but rather by multiple of such distributions or models, with some random selection mechanism at play.

1.3 Objectives

This study had the following main objective which is further broken down into specific objectives:

1.3.1 Main objective

To develop a robust Bayesian-based prognostic model of hospital outcomes for paediatrics admitted in Kenyan county hospitals.

1.3.2 Specific objectives

1. To identify research gaps and statistical limitations that are in the recently published prognostic models.
2. To explore the appropriate statistical methodology that can address the limitation identified in 1 above.
3. To use readily available routine data to develop a generic predictive model using Bayesian paradigm under the methodology identified in 2 above.

Chapter 2

Literature Review

2.1 Overview

Clinical prediction models are fundamental in medicine, they have been used to inform treatment as well as providing information on diagnostic episodes that are possibly present in a patient given his/her presenting characteristics. Prognostic models make a joint use of multiple clinical signs of a given disease and other risk prognostic factors to systematically calculate prognostic indices (probability estimates) that are used to discriminate patients of high risk from those with low risk as shown in the paper by (Mallett et al., 2010). These probabilistic discrimination is useful since it augments the clinical intuition of clinicians who have been shown by (Wakaba et al., 2014) to be in short supply especially in low-income countries. Of note, however, just a handful of models ends up being used in clinical practice out of many that are published. But this trend is not without a reason; (Mallett et al., 2010) and (Mikolajczyk et al., 2008) observed that most of the published prognostic models have been derived using poor statistical methods that adversely affect their performance during external validations.

2.2 Heterogeneity in medicine

Researchers using medical data more often than not present results for 'average' patients while ignoring potential heterogeneity of patients. This is because most of the standard statistical models, such as linear regression, estimate the population average effect, which is the

mean response of all individuals under study. But the population average effect might be a mixture of some effect, little effect and/or no effect. Thus, using customary statistical methods considerably fails to appreciate that patients are not alike! Consequently, the resultant effects of covariates on the quantity of interest tend to mislead! For instance (Schlattmann, 2009) noted that patients react differently to treatment regimes. Observable factors such as age, gender, etc. could be attributable to such differences. Owing to such heterogeneous populations, individualization of medical therapies has become necessary. However, the underlying mechanism that triggers patient variability remain unknown. Most of the factors that are thought to be attributable to patient heterogeneity are largely latent (unobserved) variables in medical research. For example, information on genetic polymorphisms of patients remain unknown yet that information is crucial in pharmacokinetic studies. Faced with unobserved patient heterogeneity, a statistical methodology that handles variability between individuals due to unobserved covariates has to be used particularly in clinical predictive models so as to improve performance.

2.3 Methodological flaws in prognosis

In clinical research, multivariate clinical information is integrated to optimally predict patient status or even disease progression. However, overwhelming evidence arising from a series of systematic reviews such as (Mallett et al., 2010; Collins et al., 2011, 2014) has pointed out numerous challenges that make prognostic models perform poorly. They include sample size, missing data and most importantly inappropriate methodological choice in their statistical analyses. Models derived from such shortcomings of methodology give overoptimistic performance during derivation. Therefore, it is not surprising that most of these suboptimal prognostic models undoubtedly never get to be used as quantified by (Collins et al., 2016). This is because they can never pass a rigorous test of external validation. Given that patient are not alike as pointed out in section 2.2, a statistical methodology that clusters patients into homogeneous subpopulations should therefore be used to account for the inherent variability in the clinical studies.

2.4 Machine Learning(ML) methods

Oermann et al. (2016) defined machine learning as a branch of Artificial Intelligence, which is an interdisciplinary field combining computer science and mathematical statistics to develop models with maximal predictive accuracy. Kourou et al. (2015) further observed that ML techniques have the ability to discover, identify patterns and relationships present in complex datasets. Oermann et al. (2016) applied ML techniques to predict individual patient outcomes after radiosurgery and reported that the approach provided the best possible predictions in relation to the methods they were compared with. Based on (Cruz and Wishart, 2006), the accuracy of cancer prediction outcome has significantly improved by 15%–20% with the application of ML techniques in the prognostic models. Clustering is a common unsupervised task of ML with the aim of determining intrinsic groupings. That is, observations with similar characteristics but dissimilar from others are grouped together to constitute subpopulations. Most of the clustering algorithms are *distance-based*; they use the geometric distance matrix as a criterion of grouping observations. The literature has shown that probability-based clustering is preferred to a distance-based. In particular, (Aitkin et al., 1981) pointed out that *"when clustering samples from a population, no cluster method is a priori believable without a statistical model"*. The *mixture model* is an example of such model-based ML methods used in clustering as we have shown below.

2.4.1 Mixture models

In statistical pattern recognition, the finite mixture model is a probabilistic clustering algorithm used to expose some natural groupings that may underlie the data (heterogeneous population). Model-based approach to clustering is supported by (Aitkin et al., 1981), who remarked in their paper that "*clustering methods based on such mixture models allow estimation and hypothesis testing within the framework of standard statistical theory.*"

The main assumption of the mixture model is that each data point y_i is a realization of the mixture density where observations in the model corresponds to clusters/subpopulations. Probabilistic (soft clustering) of the data into K clusters can be obtained in terms of the fitted posterior probabilities of component membership for the data as shown by (Schlattmann, 2009). A substantive hard clustering can be subsequently obtained by assigning each obser-

vation to the component to which it has the highest fitted posterior probability of belonging as demonstrated by (McLachlan and Peel, 1998).

2.4.2 Applications of mixture models in prognosis

- Kusmakar et al. (2016) used Gaussian Mixture Model (GMM) for the identification of psychogenic non-epileptic seizures. They reported that GMM algorithm accurately modeled the non-epileptic and epileptic movements, hence enhancing the overall predictive accuracy of the their model to 91% of the new data.

- Le and Bukkapatnam (2016) utilized a nonparametric statistical Dirichlet-Process to build a prognostic model in pulse rate dynamics with the aim of predicting real time onsets of episode before the clinical symptoms appear. They reported that the technique enhanced effective prediction.

- Lu et al. (2016) used finite mixture model to accurately segment angiography blood vessels which has played an important role in an interventional treatment of vascular diseases.

- The mixture model has been used in the identification of high-risk groups in the process of disease mapping. Schlattmann et al. (1996) used mixture model approach to identify spatial heterogeneity in tuberculosis risk and mapped it within an empirical Bayes framework.

- Gene expression profiling is expected to unveil the underlying molecular features. A parametric clustering method using the Gaussian mixture model and the Bayes inference was used by (Muro et al., 2003) and it revealed three groups of expressed genes in their study.

- Neuroblastoma patients, most of the time, experiences heterogeneous survival outcomes despite aggressive treatment. Aware of these variabilities, (Hunsberger et al., 2009) used the finite mixture model to analyze a large cohort of these patients with the aim of identifying patients with the shared types of neuroblastoma so that individualized interventional therapies could be enhanced. The robustness of these methods were tested on simulated data and misclassification rates was found to be quite low.

Chapter 3

Research Methodology

3.1 Study setting

This was a cross-sectional prospective study on patients admitted to paediatric wards across 14 county hospitals in Kenya as shown in Figure 3.1. Selection of these hospitals were purposeful to reflect high and low malaria endemicity both in large rural environments and urban settings with the condition of a given hospital admitting at least 1,000 children per year. In general, these hospitals are part of the Clinical Information Network (CIN) which is a collaboration between researchers from the Kenya Medical Research Institute (KEMRI), Wellcome Trust Research Programme (WTRP), the Kenya Ministry of Health (MOH), the Kenya Paediatric Association (KPA) and the University of Nairobi (UON). The main aim of the CIN was to use de-identified patient level routine data to improve hospital care. A detailed description of CIN study has been given elsewhere by (Ayieko et al., 2015). All research related activities undertaken by CIN has received approval from the national KEMRI scientific and ethical review committees.

3.1.1 Data management

Data were abstracted from the patient file by a trained data clerk on the daily basis following the discharge or death of a patient. The abstracted data were then fed into a customized data capture tool designed with the non-proprietary Research Electronic Data Capture (REDCap) platform which has an inbuilt range and validation checks as documented by Harris et al. (2009). Before the data were synchronized to a central database, the data clerk checked and

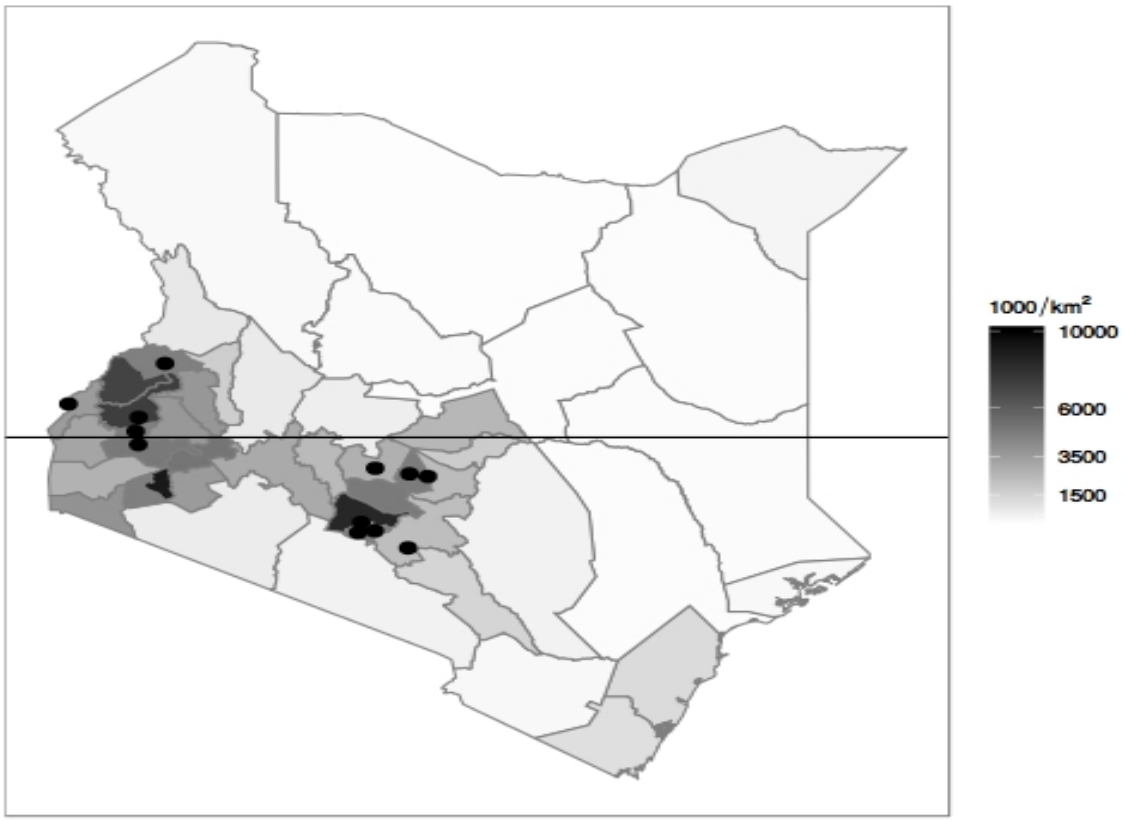

Figure 3.1: Location of hospitals under study and their population density

ensured completeness as well as consistency with locally executed cleaning scripts which were written in **R** language. If the error or any inconsistencies were detected in any record, a data clerk looked for that specific file and verified. Another round of data checks were done at the central server and in the event of any inconsistencies, data clerks whose record(s) were problematic were rang to rectify the anomaly. This loops of validation checks guaranteed a high quality data.

3.1.2 Inclusion criteria

All admissions aged $>$ 1 month hospitalized in the paeditric wards of all 14 CIN participating hospitals from September 2013 through December 2016 were deemed eligible for inclusion to the study. Patients with surgical conditions or burns were excluded because they require different clinical management. The flow of data from the server is as shown in figure 3.2 below.

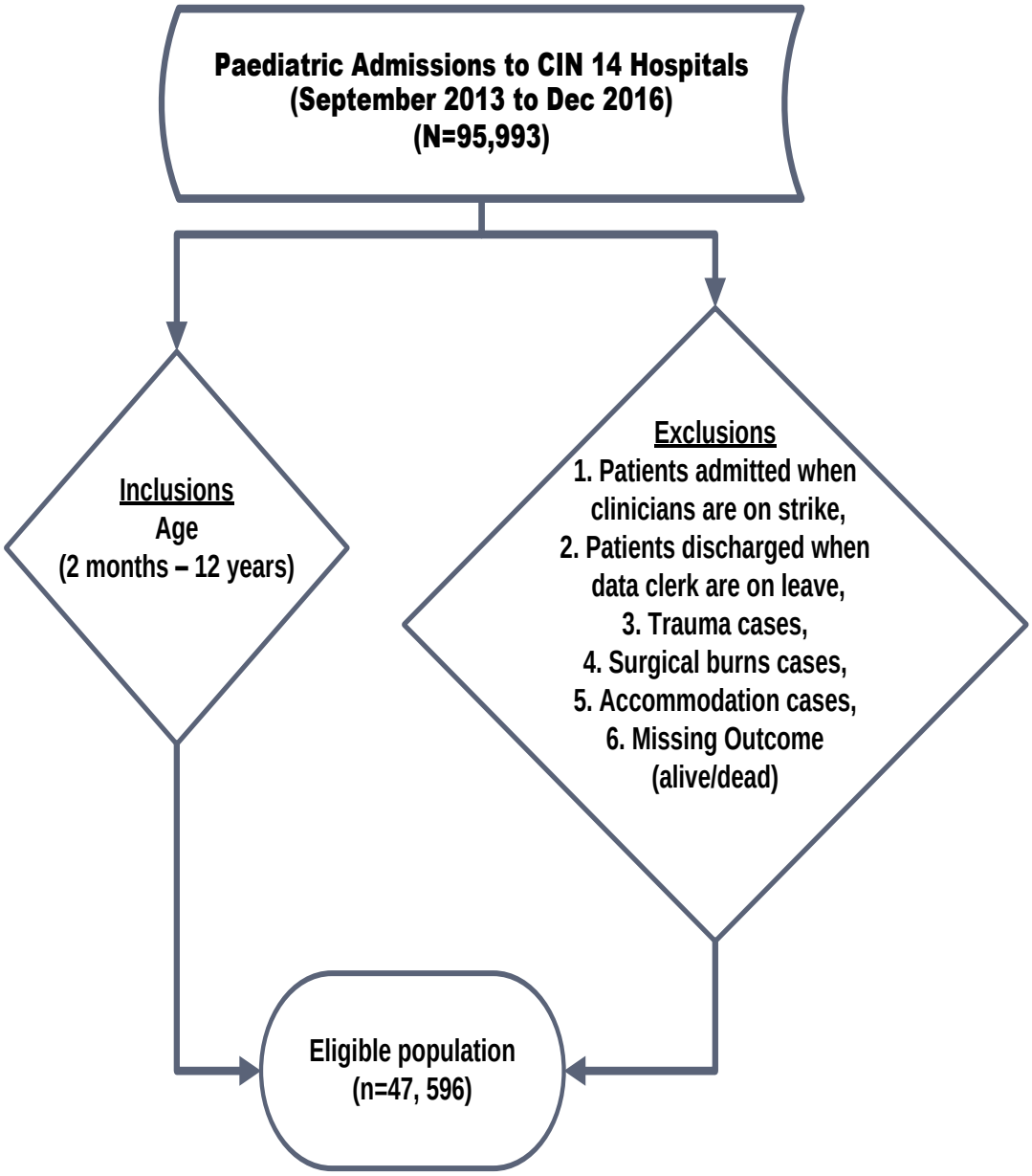

Figure 3.2: Inclusion criteria

3.2 Bayesian inference

Bayesian inference is a powerful framework that utilizes Bayes theorem to make inference. This framework is efficient because all relevant sources of uncertainty and other unknown quantities are expressed as a random variable.

In this paradigm, current information or knowledge about the parameters of the model is expressed mathematically by placing a probability distribution on the parameters, termed as *"prior distribution"*, commonly expressed as $p(\theta)$.

With the availability of the new data y, the information contained in y pertaining the model parameters is then expressed as the *"likelihood"*. By definition, the *likelihood* is directly proportional to the distribution of the observed data given the parameters of the model, expressed as $p(y \mid \theta)$.

Combination of the *likelihood* and the *prior distribution* results to a probability distribution which is updated. Commonly referred to as *"posterior distribution"*. This is where all inferences about the model are based on. Bayes' Theorem, an elementary identity in probability theory, states how the updating of posterior distribution is done mathematically as shown by (Gelman et al., 2014):

$$p(\theta \mid y) = \frac{p(\theta) \times p(y \mid \theta)}{\int_{\Theta} p(\theta) \times p(y \mid \theta)\, d\theta}.$$

Which is equivalent to,

$$p(\theta|y) \propto p(\theta) \times p(y|\theta).$$

3.2.1 Prior specification $p(\theta)$

Priors rules out unreasonable parameter values. As opposed to widely used flat priors, this study used weakly informative priors to help control inference computationally and statistically. Computationally, (Gelman et al., 2015) showed that a weakly informative prior increases the curvature around the volume where the solution is expected to lie, which in turn guides MCMC sampler by not allowing them to stray too far from the location of a surface. Statistically, weakly informative priors are not very sensitive, in the sense that reasonable changes in the prior do not produce noticeable changes in the posterior.

All models used in this study had different prior specifications based on the complexity

of a given model. Specification of each of the model is as follows;

1. **Prior Specification for finite mixture model**

   Finite Mixture model parameters has three parameters $(\boldsymbol{\theta}, \boldsymbol{\mu}, \boldsymbol{\sigma})$, where $\boldsymbol{\theta}$ is the mixture probabilities, $\boldsymbol{\mu}$ is the mixture/components means and $\boldsymbol{\sigma}$ is the standard deviation of mixtures as defined by (Gelman et al., 2014).

   The weakly informative priors of each is as defined below:

   - Dirichlet prior on the mixture proportions $p(\boldsymbol{\theta}) \sim Dirichlet(\alpha_1, \ldots, \alpha_H)$, where the sum of α_h is a measure of the strength of the prior distribution.

     The Dirichlet distribution enforces uniqueness of the components in the mixture model as proven by (Kucukelbir et al., 2015).

   - Gaussian prior on the component means

     $$p(\boldsymbol{\mu}) = \prod_{h=1}^{H} N(\mu_h; 0, \sigma_\mu).$$

   - Lognormal prior on the standard deviations of components

     $$p(\boldsymbol{\sigma}) = \prod_{h=1}^{H} logNormal(\sigma_h; 0, \sigma).$$

2. **Prior Specification for mixed effects model.**

   Mixed model parameters has four parameters $(\boldsymbol{\beta}, \boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\alpha})$,

   where $\boldsymbol{\beta}$ is the vector of model predictors, $\boldsymbol{\mu}$ is the random effect of the mixed model, $\boldsymbol{\sigma}$ is the standard deviation of random effect and $\boldsymbol{\alpha}$ is the intercept.

   Weakly informative priors of each is as defined below:

   - Half Cauchy priors on the model coefficients $p(\boldsymbol{\beta}) \sim cauchy(0, 2.5)$.
   - Cauchy priors on the model intercept $p(\boldsymbol{\alpha}) \sim cauchy(0, 10)$.
   - Gaussian prior on the random effect of the model

     $$p(\boldsymbol{\mu}) = \prod_{i=1}^{K} N(\mu_i; 0, \sigma_\mu),$$

where K is the number of random effects used in the model.

3.2.2 Sensitivity analysis on priors

Sensitivity analysis entails checks on model settings to assess the robustness of the choice of priors. Changes in priors, perhaps even slight ones often results in changes in posterior inferences especially when priors are not robust. We compared our models with plausible but different priors, different choices did not produce gross changes of inference from the posterior. We were therefore confident in the final selection of the priors.

3.2.3 Label switching

Definition of label-switching

In order to make inferences with a mixture model we had to learn each of the component weights θ_k and the component parameter λ_k. If the posterior cannot discriminate between the components, then it cannot discriminate between the component parameters. Put differently, individual component distribution is identical with other components $\pi_k(y \mid \lambda_k) = \pi(y \mid \lambda_k)$. This phenomena of permuting labels of components/subpopulations in different chains of MCMC is termed as *Label Switching*. It arises if the parameters of the mixture components have exchangeable priors $\pi(\tau(\boldsymbol{\alpha})) = \pi(\boldsymbol{\alpha})$ then it has been shown by (Gelman et al., 2014) that the posterior will inherit the permutation invariance of the mixture likelihood as follows;

$$\begin{aligned}
\pi(\sigma(\boldsymbol{\alpha}), \sigma(\boldsymbol{\lambda}) \mid y) &\propto \pi(\sigma(\boldsymbol{\alpha}))\, \pi(\sigma(\boldsymbol{\lambda})) \sum_{k=1}^{K} \lambda_{\sigma(k)}\, \pi_{\sigma(k)}(y \mid \alpha_{\sigma(k)}), \\
&\propto \pi(\sigma(\boldsymbol{\alpha})),\, \pi(\sigma(\boldsymbol{\lambda})) \sum_{k'=1}^{K} \lambda_{k'}\, \pi_{k'}(y \mid \alpha_{k'}), \\
&\propto \pi(\boldsymbol{\alpha})\, \pi(\boldsymbol{\lambda}) \sum_{k'=1}^{K} \lambda_{k'}\, \pi_{k'}(y \mid \alpha_{k'}), \\
&= \pi(\boldsymbol{\alpha}, \boldsymbol{\lambda} \mid y).
\end{aligned}$$

Hence it becomes ambiguous as to which parameters λ are associated with each component in the mixture.

3.2.4 Label-switching solutions

Ordering of constraints

To avoid the phenomena of label-switching in the mixture model we made exchangeable prior non-exchangeable by imposing an ordering on parameter space. This technique has been shown by (Gelman et al., 2014) to be an efficient computational trick which does not affect the resulting inferences. The solution is defined as follows: given the exchangeable priors $\pi(\boldsymbol{\alpha})$ non-exchangeable prior can be enforced by ordering as shown in **equation 3.1**

$$\pi'(\boldsymbol{\alpha}) = \begin{cases} \pi(\boldsymbol{\alpha}), & \alpha_1 \leq \ldots \leq \alpha_K \\ 0, & \text{else} \end{cases} \tag{3.1}$$

Further details how ordering of parameters works in a mixture model is demonstrated in **section 3.2.5**

3.2.5 Ordering in a Two-component case

The following computations shows how ordering works in a two-mixture components as it was the case in this study. In the two-component case there are two parameters α_1 and α_2 and mixture weights θ_1 and $\theta_2 = 1 - \theta_1$ The desired expectation can be decomposed over two pyramids that are present in a two-dimensional parameter space as follows;

$$\begin{aligned}
\mathbb{E}_\pi[f] &= \int \mathrm{d}\theta_1 \mathrm{d}\theta_2 \mathrm{d}\alpha_1 \mathrm{d}\alpha_2 \cdot f(\alpha_1, \alpha_2) \cdot \pi(\alpha_1, \alpha_2, \theta_1, \theta_2 \mid y) \\
&\propto \int \mathrm{d}\theta_1 \mathrm{d}\theta_2 \mathrm{d}\alpha_1 \mathrm{d}\alpha_2 \cdot f(\alpha_1, \alpha_2) \cdot \pi(\alpha_1, \alpha_2)\, \pi(\theta_1, \theta_2)\, (\theta_1 \pi(y \mid \alpha_1) + \theta_2 \pi(y \mid \alpha_2)) \\
&\propto \int_{\alpha_1 < \alpha_2} \mathrm{d}\theta_1 \mathrm{d}\theta_2 \mathrm{d}\alpha_1 \mathrm{d}\alpha_2 \cdot f(\alpha_1, \alpha_2) \cdot \pi(\alpha_1, \alpha_2)\, \pi(\theta_1, \theta_2)\, (\theta_1 \pi(y \mid \alpha_1) + \theta_2 \pi(y \mid \alpha_2)) \\
&\quad + \int_{\alpha_2 < \alpha_1} \mathrm{d}\theta_1 \mathrm{d}\theta_2 \mathrm{d}\alpha_1 \mathrm{d}\alpha_2 \cdot f(\alpha_1, \alpha_2) \cdot \pi(\alpha_1, \alpha_2)\, \pi(\theta_1, \theta_2)\, (\theta_1 \pi(y \mid \alpha_1) + \theta_2 \pi(y \mid \alpha_2)).
\end{aligned}$$

If we permute parameters as

$$(\alpha_1, \alpha_2) \to (\beta_2, \beta_1)$$

and

$$(\theta_1, \theta_2) \to (\lambda_2, \lambda_1),$$

then the second pyramid will be rotated into first as follows;

$$\begin{aligned}
\mathbb{E}_\pi[f] \propto & \int_{\alpha_1<\alpha_2} \mathrm{d}\theta_1\mathrm{d}\theta_2\mathrm{d}\alpha_1\mathrm{d}\alpha_2 \cdot f(\alpha_1,\alpha_2)\cdot \pi(\alpha_1,\alpha_2)\,\pi(\theta_1,\theta_2)\,(\theta_1\pi(y\mid\alpha_1)+\theta_2\pi(y\mid\alpha_2)) \\
& + \int_{\beta_1<\beta_2} \mathrm{d}\lambda_2\mathrm{d}\lambda_1\mathrm{d}\beta_2\mathrm{d}\beta_1 \cdot f(\beta_2,\beta_1)\cdot \pi(\beta_2,\beta_1)\,\pi(\lambda_2,\lambda_1)\,(\lambda_2\pi(y\mid\beta_2)+\lambda_1\pi(y\mid\beta_1)) \\
\propto & \int_{\alpha_1<\alpha_2} \mathrm{d}\theta_1\mathrm{d}\theta_2\mathrm{d}\alpha_1\mathrm{d}\alpha_2 \cdot f(\alpha_1,\alpha_2)\cdot \pi(\alpha_1,\alpha_2)\,\pi(\theta_1,\theta_2)\,(\theta_1\pi(y\mid\alpha_1)+\theta_2\pi(y\mid\alpha_2)) \\
& + \int_{\beta_1<\beta_2} \mathrm{d}\lambda_1\mathrm{d}\lambda_2\mathrm{d}\beta_1\mathrm{d}\beta_2 \cdot f(\beta_2,\beta_1)\cdot \pi(\beta_2,\beta_1)\,\pi(\lambda_2,\lambda_1)\,(\lambda_1\pi(y\mid\beta_1)+\lambda_2\pi(y\mid\beta_2)).
\end{aligned}$$

Applying permutation-invariance of f and the exchangeability of the priors the second term will be equivalent to the first term as follows:

$$\begin{aligned}
\mathbb{E}_\pi[f] \propto & \int_{\alpha_1<\alpha_2} \mathrm{d}\theta_1\mathrm{d}\theta_2\mathrm{d}\alpha_1\mathrm{d}\alpha_2 \cdot f(\alpha_1,\alpha_2)\cdot \pi(\alpha_1,\alpha_2)\,\pi(\theta_1,\theta_2)\,(\theta_1\pi(y\mid\alpha_1)+\theta_2\pi(y\mid\alpha_2)) \\
& + \int_{\beta_1<\beta_2} \mathrm{d}\lambda_1\mathrm{d}\lambda_2\mathrm{d}\beta_1\mathrm{d}\beta_2 \cdot f(\beta_1,\beta_2)\cdot \pi(\beta_1,\beta_2)\,\pi(\lambda_1,\lambda_2)\,(\lambda_1\pi(y\mid\beta_1)+\lambda_2\pi(y\mid\beta_2)) \\
\propto & \, 2\int_{\alpha_1<\alpha_2} \mathrm{d}\theta_1\mathrm{d}\theta_2\mathrm{d}\alpha_1\mathrm{d}\alpha_2 \cdot f(\alpha_1,\alpha_2)\cdot \pi(\alpha_1,\alpha_2)\,\pi(\theta_1,\theta_2)\,(\theta_1\pi(y\mid\alpha_1)+\theta_2\pi(y\mid\alpha_2)) \\
\propto & \int_{\alpha_1<\alpha_2} \mathrm{d}\theta_1\mathrm{d}\theta_2\mathrm{d}\alpha_1\mathrm{d}\alpha_2 \cdot f(\alpha_1,\alpha_2)\cdot 2\pi(\alpha_1,\alpha_2)\,\pi(\theta_1,\theta_2)\,2(\theta_1\pi(y\mid\alpha_1)+\theta_2\pi(y\mid\alpha_2)) \\
= & \int_{\alpha_1<\alpha_2} \mathrm{d}\theta_1\mathrm{d}\theta_2\mathrm{d}\alpha_1\mathrm{d}\alpha_2 \cdot f(\alpha_1,\alpha_2)\cdot \pi'(\alpha_1,\alpha_2,\theta_1,\theta_2\mid y).
\end{aligned}$$

Thus, the first and the second terms can be expressed as;

$$\mathbb{E}_\pi[f] = \mathbb{E}_{\pi'}[f]. \tag{3.2}$$

Equation **3.1** is a simplified form of equation **3.2**. Hence it has been shown that taking expectation over the pyramid defined by the standard ordering yields the same value as the expectation taken over the entire parameter space.

3.2.6 Computations

All models in this study were implemented in *Stan*. This is a flexible probabilistic programming framework originally designed by (Gelman et al., 2015) for sampling-based inference. We had strong reasons for using *Stan* and they included:

- *Stan* uses a No-U-Turn (NUT) sampler which is a variant of Hamiltonian Monte Carlo

(HMC). This sampler technically overcomes a local random-walk behavior which is a common problem in Gibbs sampler and Metropolis algorithms as excellently shown in (Betancourt, 2013). These inefficiencies costs a lot of time zigging and zagging while traversing through the target distribution. As demonstrated by (Gelman et al., 2014), such random-walk behaviors are very common when BUGS/JUGS is used to estimate complicated models such as mixture models.

- *Stan* is coded in C++ which implies that computational time is relatively short as compared to BUGS/JUGS.

- *Stan* implements Automatc Differentiation Variational Inference (ADVI) which is essential in the computation of machine learning (mixture models) posteriors which are otherwise computationally intractable because of their closed form. ADVI as documented by (Kucukelbir et al., 2015) transforms the joint density of any differentiable probability model to the real coordinate space.

3.2.7 Implementing Bayesian finite mixture model

Let $z \in \{0, \ldots, K\}$ be an assignment that indicates to which data generating process a given sample are generated from. Suppose each of the $y = y_1, ..., y_n$ items in the sample belong to one of H subpopulations. For Mathematical convenience, mixture models are often formulated in terms of latent variables indicator z_{ih}. Defined as follows

$$z_{ih} = \begin{cases} 1 & \text{if } i^{th} \text{ unit is drawn from } h^{th} \text{ component} \\ 0, & \text{otherwise} \end{cases}$$

Latent variable indicator are never observed and their values are never known beforehand. Let $h = 1, ..., H$, then the h^{th} component distribution is defined as $f_h(y_i \mid \lambda_h)$ which depends on parameter vector λ_h.

Let θ_h denote the mixing weight of the population from the component h, such that;

$$\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K),\ 0 \leq \theta_k \leq 1,\ \sum_{h=1}^{H} \theta_h = 1,$$

the joint likelihood over observations and its subpopulation becomes,

$$\pi(y, z \mid \boldsymbol{\alpha}, \boldsymbol{\theta}) = \pi(y \mid \boldsymbol{\alpha}, z)\, \pi(z \mid \boldsymbol{\lambda}) = \pi_z(y \mid \alpha_z)\, \theta_z.$$

But, in practice assignment of observation to a subpopulation uses discrete parameters which are difficult to fit accurately because gradient computations are not tractable. As a result, inference of such parameters can be done via marginalization of the assignments which yields a likelihood that depends on only continuous parameters which is essentially a convex combination of the data generating process as follows:

$$\begin{aligned}
\pi(y \mid \boldsymbol{\alpha}, \boldsymbol{\theta}) &= \sum_z \pi(y, z \mid \boldsymbol{\alpha}, \boldsymbol{\theta}) \\
&= \sum_z \pi_z(y \mid \alpha_z)\, \theta_z \\
&= \sum_{k=1}^{K} \pi_k(y \mid \alpha_k)\, \theta_k \\
&= \sum_{k=1}^{K} \lambda_k\, \pi_k(y \mid \alpha_k).
\end{aligned}$$

The sampling distribution of y then becomes;

$$p(y_i \mid \lambda, \theta) = \theta_1 f(y_i \mid \lambda_1) + \theta_2 f(y_i \mid \lambda_2) + ... + \theta_H f(y_i \mid \lambda_H). \tag{3.3}$$

To incorporate the latent variable indicator $z_i = (z_{i1}, ..., z_{iH})$ is modeled as multinomial with parameter θ. Mathematically defined as:

$$z \sim Multinormial(\theta),$$

where $\theta = (\theta_1, ..., \theta_H)^T$ is the vector of probability also known as mixing weights. The generative process of each observation is then defined as shown below

$$y \mid z = h \sim Gaussian(\mu_k, \sigma_k), \tag{3.4}$$

where h is a component(subpopulation) in the mixture model.
Hence the joint distribution of observed y and a latent indicator z on the model parameter

can be expressed as follows:

$$p(y, z \mid \lambda, \theta) = p(z \mid \theta)p(y \mid z, \lambda) = \prod_{i=1}^{n} \prod_{h=1}^{H} (\theta_h f(y_i \mid \lambda_h))^{z_{ih}}. \tag{3.5}$$

3.2.8 Bayesian mixture model posteriors

Bayesian inference over a mixture model requires both mixture likelihood as well as prior distributions for both the component parameters, λ, and the mixture probabilities, θ. The posterior distribution for multiple N observations becomes

$$\pi(\boldsymbol{\lambda}, \boldsymbol{\theta} \mid \mathbf{y}) \propto \pi(\boldsymbol{\lambda})\, \pi(\boldsymbol{\theta}) \sum_{n=1}^{N} \sum_{h=1}^{H} \theta_H\, \pi_H(y_n \mid \lambda_H), \tag{3.6}$$

where H is the number of subpopulations, $\boldsymbol{\theta}$ is the vector of mixing weights and $\boldsymbol{\lambda}$ is the vector of the model parameters.

3.3 Analysis

3.3.1 Handling missing data

Before fitting regression models we explored the levels of missingness both in explanatory and dependent variables. Since the degree of missingness was not ignorable ($< 1\%$), we conducted multiple imputation under the assumption of Missing At Random (MAR) missingness mechanism. Using chained equations, we generated 25 dataset with 100 iterations each. Rubin's rules were used to pool estimates of the model from each of the dataset imputed.

3.3.2 Analysis plan

Determining prognostic factors

Opinion from experts (pediatricians) and literature search informed the list of variables that were deemed to be of clinical relevance in predicting mortality of paediatrics. They included:

- Unconsciouness on AVPU scale
- Central Cyanosis
- Grunting
- Severe pallor
- Acidotic Breathing
- Inability to drink/breasfeed
- History of fever
- History of diarrhoea
- Indrawing
- Pallor
- Stiff Neck

- Skin Pinch
- Jaundice
- Oedema
- Number of severe comorbidities
- Time(seconds) to capillary refill
- Pulse(Normal or weak)
- Gender

Variables listed above have been shown to be significant factors of mortality by (Gathara et al., 2017).

Variable selection

Final prognostic model included all of the above variables which were selected *a prior* based on the clinical relevance. As a result, no model building process ensued since these variables were independent predictors of mortality. Statistical procedures (e.g. stepwise selection) with a strong focus on significance levels to include or exclude a variable have been highly discouraged by (Wynants et al., 2016) and (Ogundimu et al., 2016) who observed that models built from variable selection procedures such stepwise in a binary outcome tends to have a considerable bias in their coefficients especially in small sample sizes.

Determining the number of subpopulations, K

The first (Model 1) logistic mixed effect model with hospital identity as clustering factors was fitted with all the listed variables as follows:

Let the linear predictor η be the combination of the fixed effects (Prognostic factors listed above) and random effects (Hospital identity) excluding the residuals. Our main outcome was mortality(alive/dead) a binary outcome. As a result we used a logistic link function

$g(\cdot)$.

$$\boldsymbol{\eta} = \boldsymbol{X\beta} + \boldsymbol{Z\gamma},$$
$$g(\cdot) = \text{link function},$$
$$g(\cdot) = log_e(\tfrac{p}{1-p}),$$
$$h(\cdot) = g^{-1}(\cdot) = \text{inverse link function},$$

where the conditional expectation of y is given by $g(E(\mathbf{y})) = \boldsymbol{\eta}$,

$\boldsymbol{X}$ is a $\boldsymbol{N} \times p$ matrix of the p predictor variables; $\boldsymbol{\beta}$ is a $p \times 1$ column vector of the fixed-effects regression coefficients; z is the $\boldsymbol{N} \times q$ design matrix for the q random effects; γ is a $q \times 1$ vector of the random effects; and the linear predictors of the model consist of $\boldsymbol{X\beta}$.

Using kernel density estimators of the linear predictors, the number of components $\boldsymbol{K}$ was determined by visually inspecting the plot.

Mixture model specification

Bayesian Mixture Model was specified using an empirical prior distribution for the mean of the K normal distributions.

$$\pi(y_1, \ldots, y_N \mid \mu_1, \sigma_1, ..., \mu_k, \sigma_k, \theta_1, \ldots, \theta_k) = \sum_{n=1}^{N} \theta_1 \mathcal{N}(y_n \mid \mu_1, \sigma_1) + \ldots + \theta_k \mathcal{N}(y_n \mid \mu_k, \sigma_k),$$

where $\boldsymbol{K}$ is the number of components which was obtained as explained above.

Mixture model assumption: Mixture model assumes that each observation y_i is drawn from one of the components $\boldsymbol{K}$ independently as opposed to the entire dataset being drawn from one of the components. Figure 3.3 is the graphical model where θ is the vector of mixing parameters, μ is the vector of means for the component distributions, α is the prior mean, and σ represents the prior standard deviation.

3.3.3 Hierarchical modeling

This is a statistical model with either parameters or data structured at different clustering levels. In most cases, though not necessarily in all cases, data are nested within each other (Gelman et al., 2014). These models consist both fixed effects and random effect (also known as clustering factor). To derive the final prognostic model, hospital identity and the number

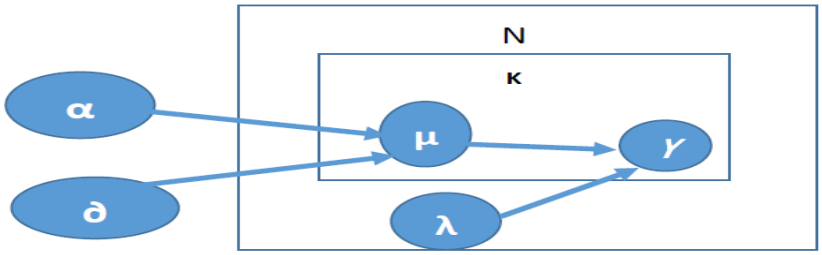

Figure 3.3: Bayesian graphical model

of subpopulations were used as a random effect.

3.3.4 Procedure of sampling

A set of 4 chains each with 2000 iterations for the mixture model and 1000 iterations for each of the two mixed models were fitted using Stan. To have a good starting point for Markov chain Monte Carlo (MCMC) runs, we threw away(burn-in) some iterations at the beginning of the run for different models as follows; first 200 iterations for mixture model and 100 iterations for each of the two mixed models were discarded as burn-in in each sample. To reduce autocorrelation in the estimates of the model, a thinning interval of 10 was used.

3.3.5 Assessing diagnostics of MCMC

Convergence of the Bayesian models was assessed by visually inspecting their trace-plots. Convergence was assumed if the chains intermingled without a definite pattern. Furthermore, we used a scale reduction statistic $\hat{R}$ (R-hat) as recommended by (Gelman and Rubin, 1992). Ideally, $\hat{R}$ statistic quantifies the ratio of variance of the draws pooled from all chains to the average variance of the draws for each chain. Therefore, at equilibrium state all chains ought to have a $\hat{R}$ of 1. If any chain has not converged to a common target distribution the $\hat{R}$ will be greater than 1.

3.3.6 Posterior Predictive Distribution (PPD) of the model

To assess whether our model could be used to make predictions for the out-of-sample data, we used posterior predictive density which is defined as follows. Given observation data y and a model parameters θ, the posterior predictive distribution for new observation $\tilde{y}$ is

$$p(\tilde{y} \mid y) \;=\; \int_{\Theta} p(\tilde{y} \mid \theta)\; p(\theta \mid y)\; \mathrm{d}\theta,$$

and the log posterior predictive density defined as;

$$\log p(\tilde{y} \mid y) = \log \int_{\Theta} p(\tilde{y} \mid \theta)\; p(\theta \mid y)\; \mathrm{d}\theta.$$

Thus to validate the fit of our model, we generated new simulated data from the posterior density and calculated the test statistics such as mean, standard deviation, minimum and maximum which were compared visually with the observed data as recommended by (Gelman et al., 2014).

3.3.7 Cross validation and model comparison

We used the Watanabe-Akaike information criterion (WAIC) and Leave-one-out cross-validation (LOO) methods for Bayesian models as recommended by (Gelman et al., 2014) and (Vehtari et al., 2016) to perform an internal validation of the derived model. To justify the added complexity in the derived model, we compared its predictive performance against a model without a random effect of the subpopulation. These methods uses posterior simulations to estimate the out-of-sample predictive value.

Chapter 4

Results and Findings

4.1 Study population

4.1.1 Baseline characteristics of derivation cohort

This study included 47,596 participants who met eligibility criteria shown in Figure 3.2. Proportion of female participants was 44.93% and age in months had a median of 20 (Range 10-46). Mortality varied considerably between hospitals with overall mortality being 2,399 (5.04%) with a range of 0% to 8.8% as shown in Figure 4.1. More than half of the study population (76.1%) had a history of fever which suggested a possibility of malaria parasitemia amongst the admissions. A detailed distribution of key indicators are summarized in Table 4.1.

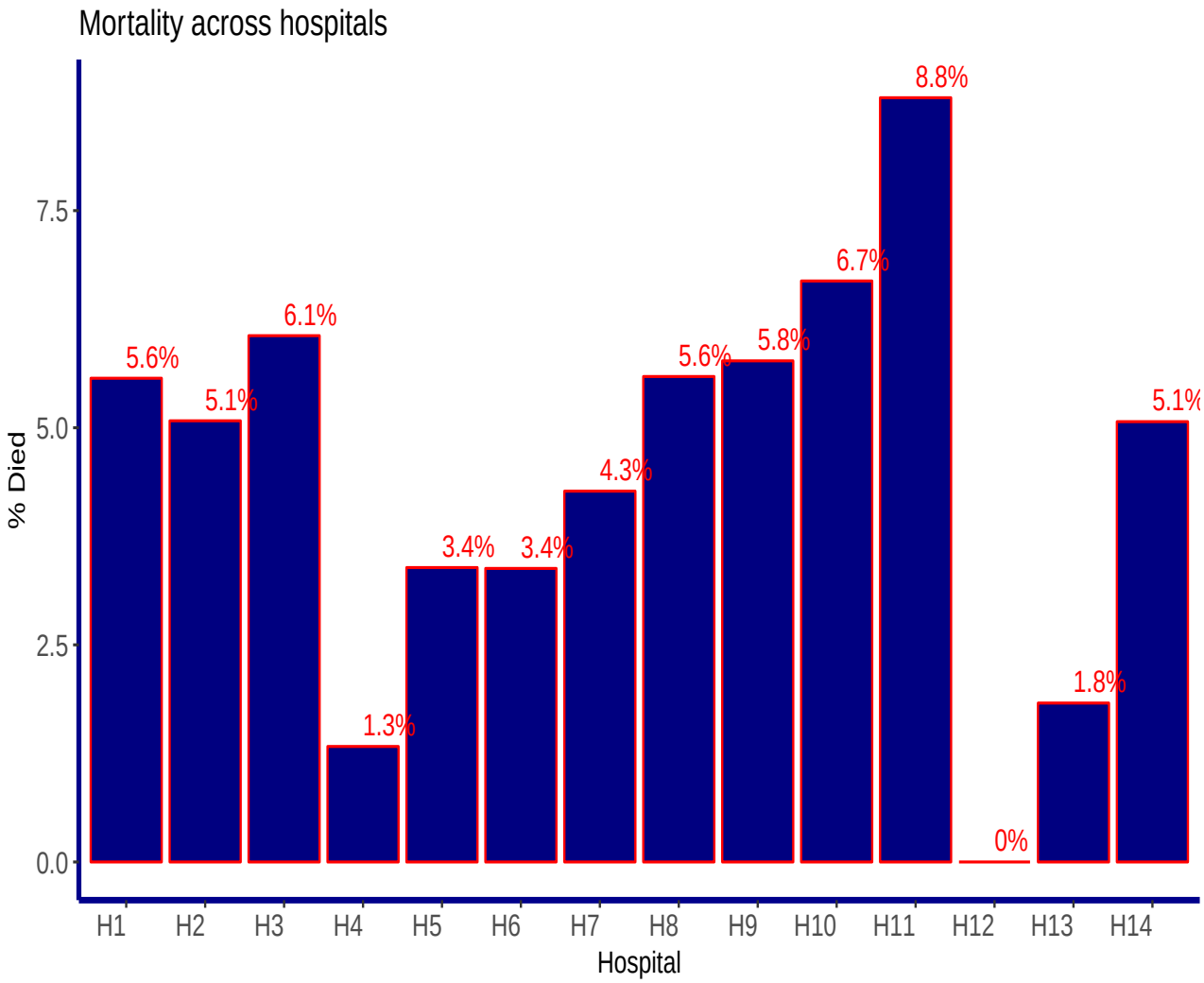

Figure 4.1: Distribution of mortality in study hospitals

Table 4.1: Population characteristics

Indicator	Count (%)
Cases	47596
Mortality	2399 (5.04)
Female	21384 (44.93)
Median age in Months(IQR)	20 (10 - 46)
Median weight(kgs) (IQR)	10 (7.5 - 14)
Prostration	3881 (8.15)
Grunting	4884 (10.26)
Can't Drink	7535 (15.83)
History of Fever	36219 (76.1)
History of Diarrhea	14917 (31.34)
Convulsion	10462 (22.1)
AVPU<A	3045 (6.4)
Severe pallor	2895 (6.08)
Central cyanosis	278 (0.58)
Respiratory distress	16115 (33.86)
Impaired Consciousness	4274 (8.98)

4.2 Prognostic model derivation

4.2.1 Detecting existence of subgroups in the larger population

Using Kernel density plots of the linear predictors expressed as

$$\eta_i = \alpha + \beta_1 \text{AVPU}_i + \beta_2 \text{Central Cynosis}_i + \ldots + \beta_p \text{GenderMale}_i$$

of Model 1, two homogeneous subgroups were identified in the study population as shown in Figure 4.2. Visual inspection of the plot suggested a possibility of two data generating mechanisms and also the existence of latent variables responsible in distinguishing two subgroups. Hence Model 1 was limited in this respect since it couldn't account for such latent variables.

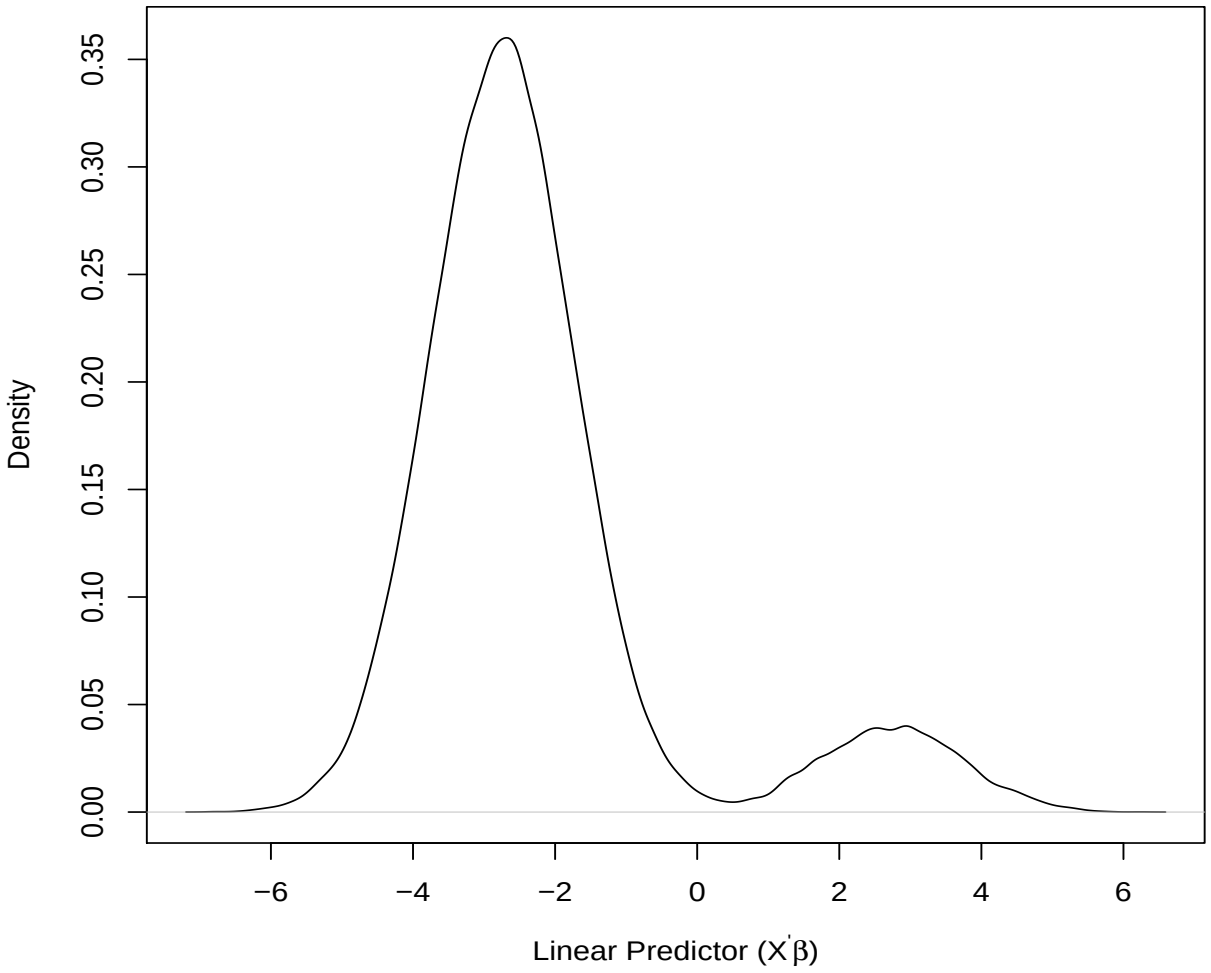

Figure 4.2: Subpopulations in the study population

4.2.2 Identifying patients in each of the subpopulation using mixture models

With the assumption that each observation y_i could be drawn from one of 2 data generating processes as shown in Figure 4.2, we developed a mixture model where we conditioned on the assignment $z \in \{1, 2\}$ to indicate to which data generating process each of 47,596 observations used in this study were generated.

Convergence of the mixture model was assessed using trace-plots shown in Figure 4.3 and we were convinced that all 4 chains converged to the target distribution. The $\hat{R}$ was approximately 1 as shown in Table 4.2 which further suggested a convergence to equilibrium distribution. The dreaded *label switching* which is an inherent problem in mixure model was not of a concern in this case as shown in 4.4. From the results of the estimates of the parameters as shown in Table 4.2, the weights of the two mixture components were approximately $\theta_1 = 0.809$ and $\theta_2 = 0.191$.

Table 4.2: Mixture model estimates

Parameter	Mean	standard deviation	95% Credible Interval	Rhat
θ_1	0.809	0.008	0.793 - 0.825	1.00
θ_2	0.191	0.008	0.175 - 0.207	1.00
σ_1	0.735	0.005	0.724 - 0.745	1.00
σ_2	1.298	0.005	1.267 - 1.331	1.00
μ_1	-3.795	0.007	-3.808 - -3.782	1.00
μ_2	-2.287	0.049	-2.379 - -2.188	1.00

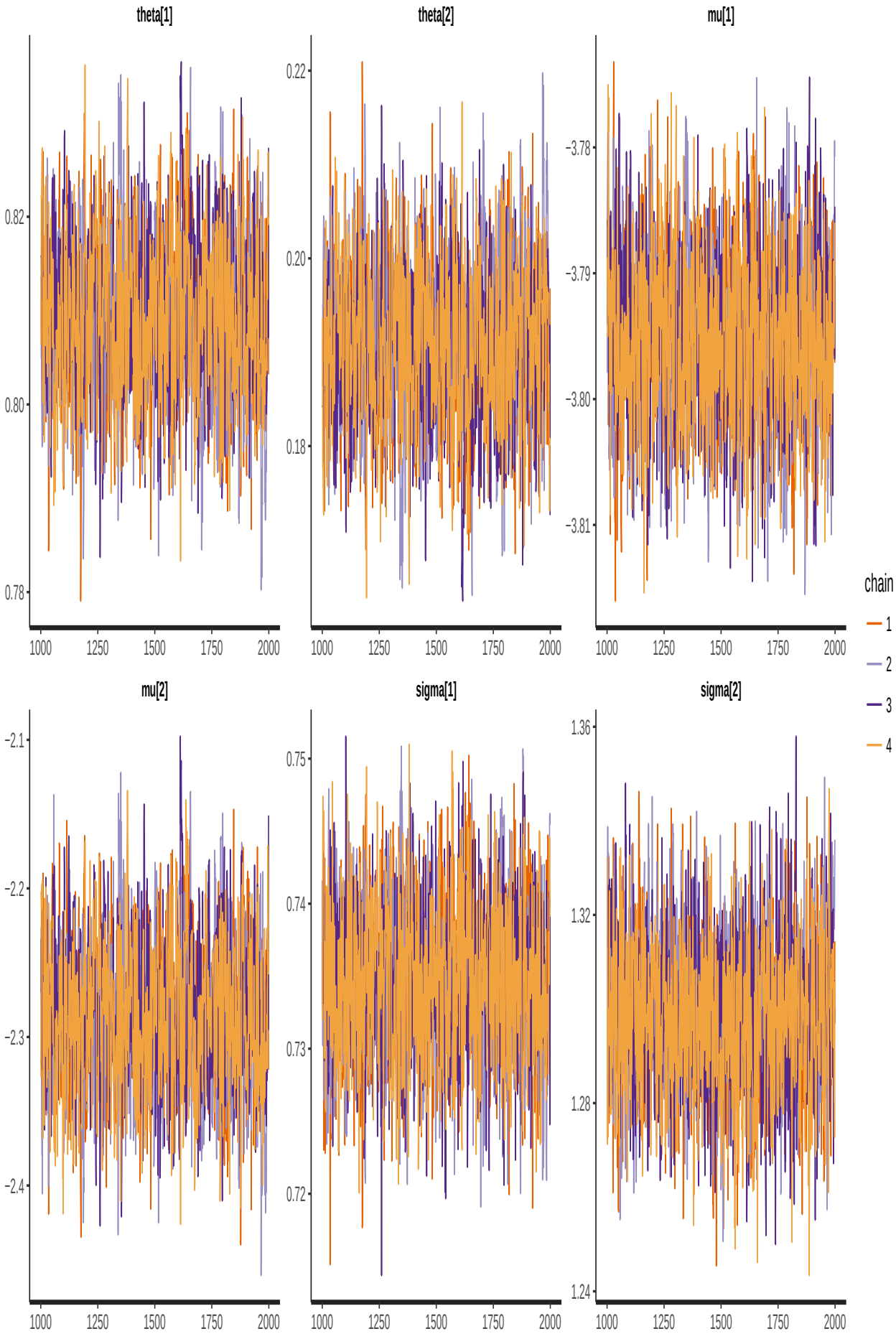

Figure 4.3: Convergence of mixture model

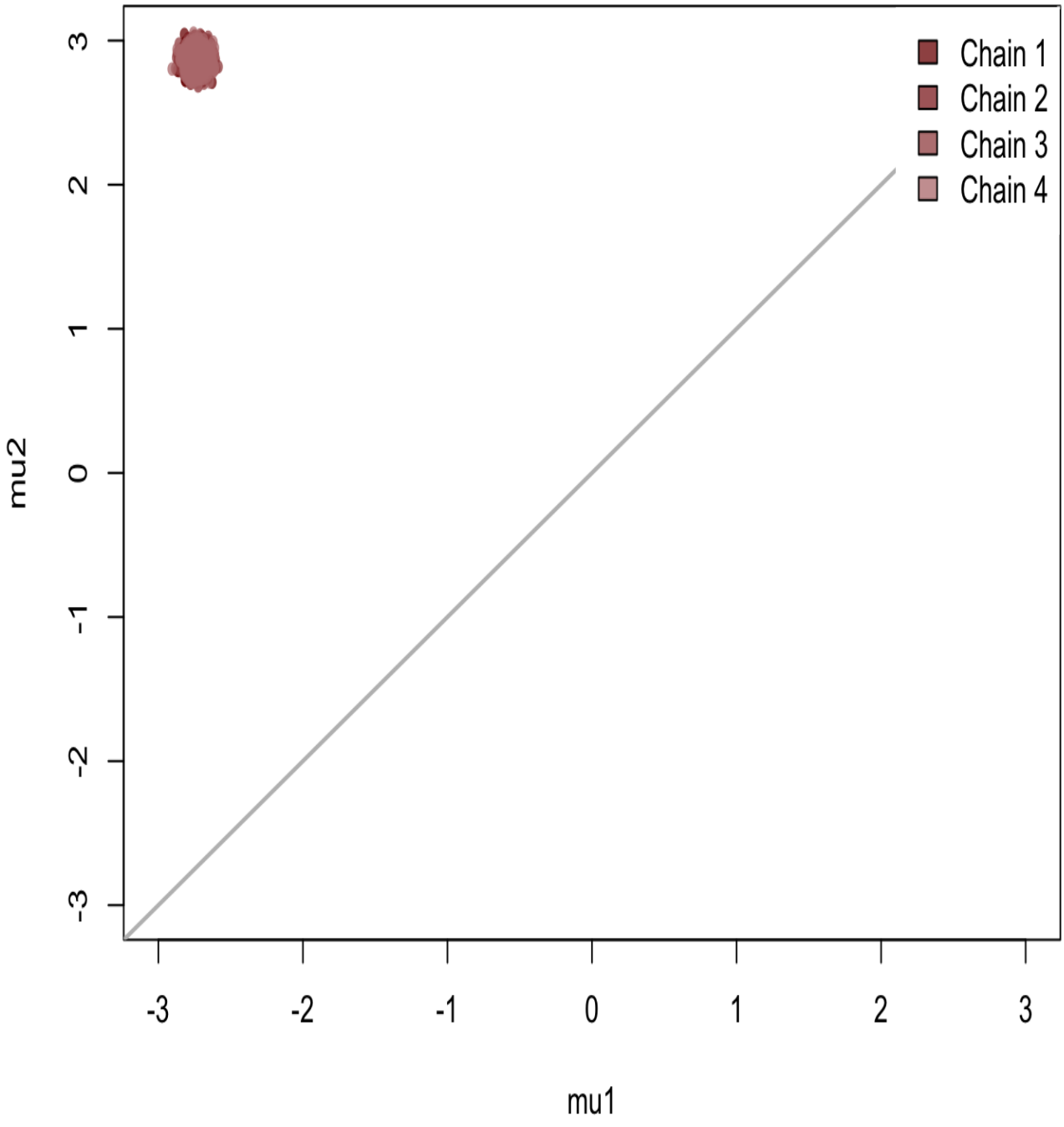

Figure 4.4: Label-switching diagnostics

Data Generating Processes

So as to make predictions, probabilistic clustering (soft clustering) were obtained from the fitted posterior probabilities of component membership. For each 47,596 observations, we picked one (hard clustering) of the two mixture components which had the highest probability and its corresponding probability. The distribution of each of probability for each of the component is as shown in Figure 4.5

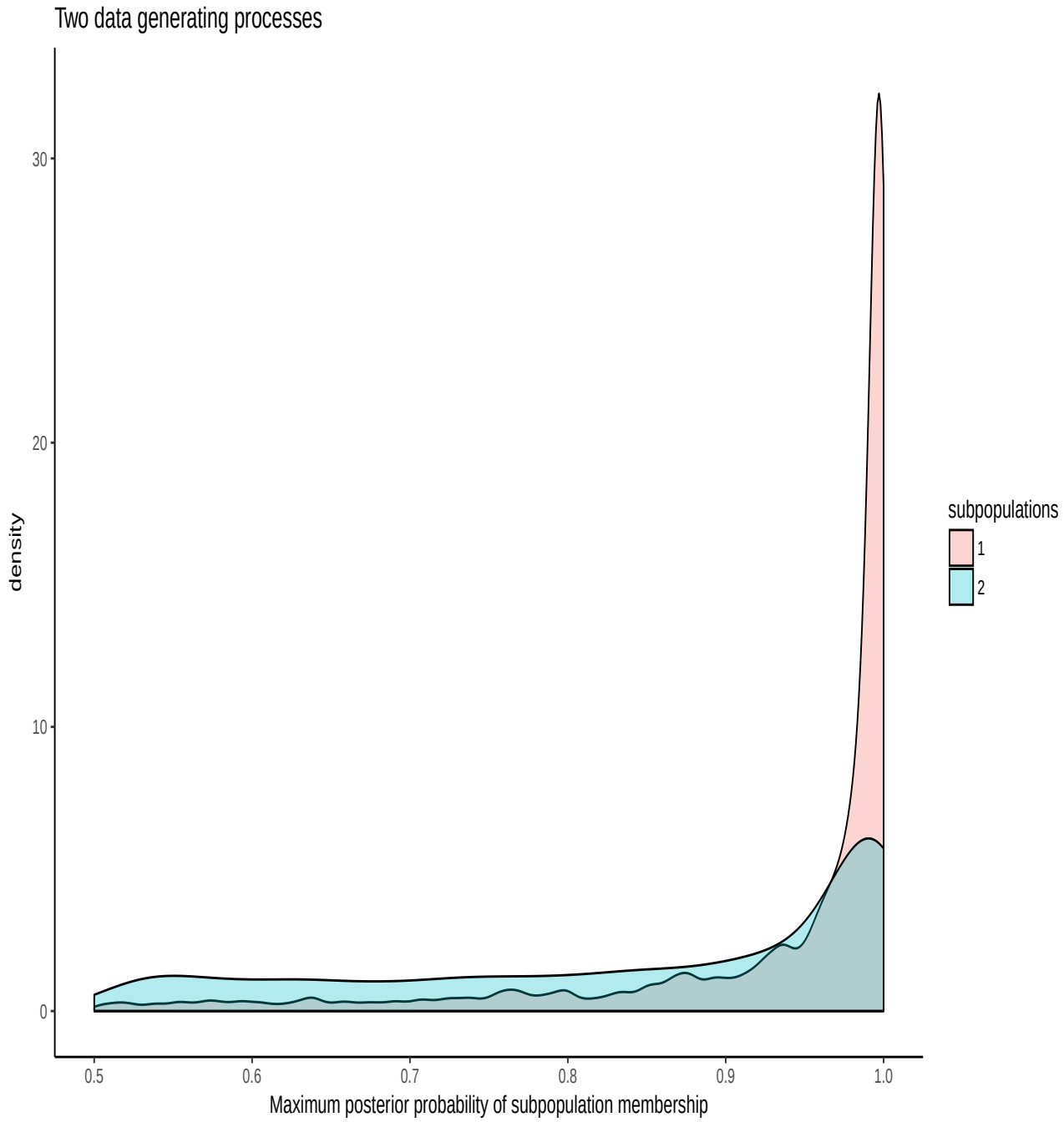

Figure 4.5: Data generating processes

4.2.3 Prognostic model results

All 4 chains used in estimating the prognostic logistic model converged to the target distribution in all of the predictors we were considering as shown in Figure 4.6. As a result, we proceeded to make inference. Odds ratios (ORs) and 95 per cent confidence intervals (CIs) are included to enhance interpretability. The AVPU scale which measures unconsciousness was the strongest predictor of mortality with odds of (AOR=2.94, 95% CI= 2.57 - 3.36). Oedema (AOR= 2.66, 95% CI= 2.18 - 3.24) , pallor (AOR=2.09, 95% CI= 1.86 - 2.36) and the presence of >= 3 severe comorbidities (AOR=2.19, 95% CI= 1.73 - 2.74) were also associated with an increased risk of death as shown in Table 4.3 and in Figure 4.7. Indrawing was the single most respiratory distress that the strongest predictor of mortality.

Table 4.3: Prognostic model estimates

Predictor	Adjusted Coefficient(Odds ratio)	95% Credible Interval
AVPU Not Alert	2.94	2.57 - 3.36
central cyanosis	1.88	1.32 - 2.67
Grunting	0.84	0.71 - 0.99
Acidiotic breath	1.70	1.42 - 2.05
Severe pallor	0.84	0.71 - 0.99
Acidiotic breath	1.70	1.42 - 2.05
Cannot drink	1.89	1.7 - 2.11
History of fever	0.82	0.74 - 0.91
History of diarrhoea	1.44	1.3 - 1.59
Indrawing	2.04	1.83 - 2.28
Pallor	2.09	1.86 - 2.36
StiffNeck	1.97	1.6 - 2.42
skin pinch 1-2 seconds	1.23	1.09 - 1.38
Skin pinch >= 2 seconds	1.91	1.62 - 2.23
Jaundice	1.28	0.99 - 1.62
Oedema	2.66	2.18 - 3.24
1 comorbidity	1.10	0.9 - 1.33
2 comorbidity	1.75	1.43 - 2.15
>= 3 comorbidity	2.19	1.73 - 2.74
capillary refill 2 seconds	1.04	0.93 - 1.16
capillary refill Indeterminate	0.84	0.62 - 1.11
capillary refill > 2seconds	1.31	1.12 - 1.53
Weak Pulse	1.74	1.51 - 2.00
Male	0.80	0.73 - 0.88

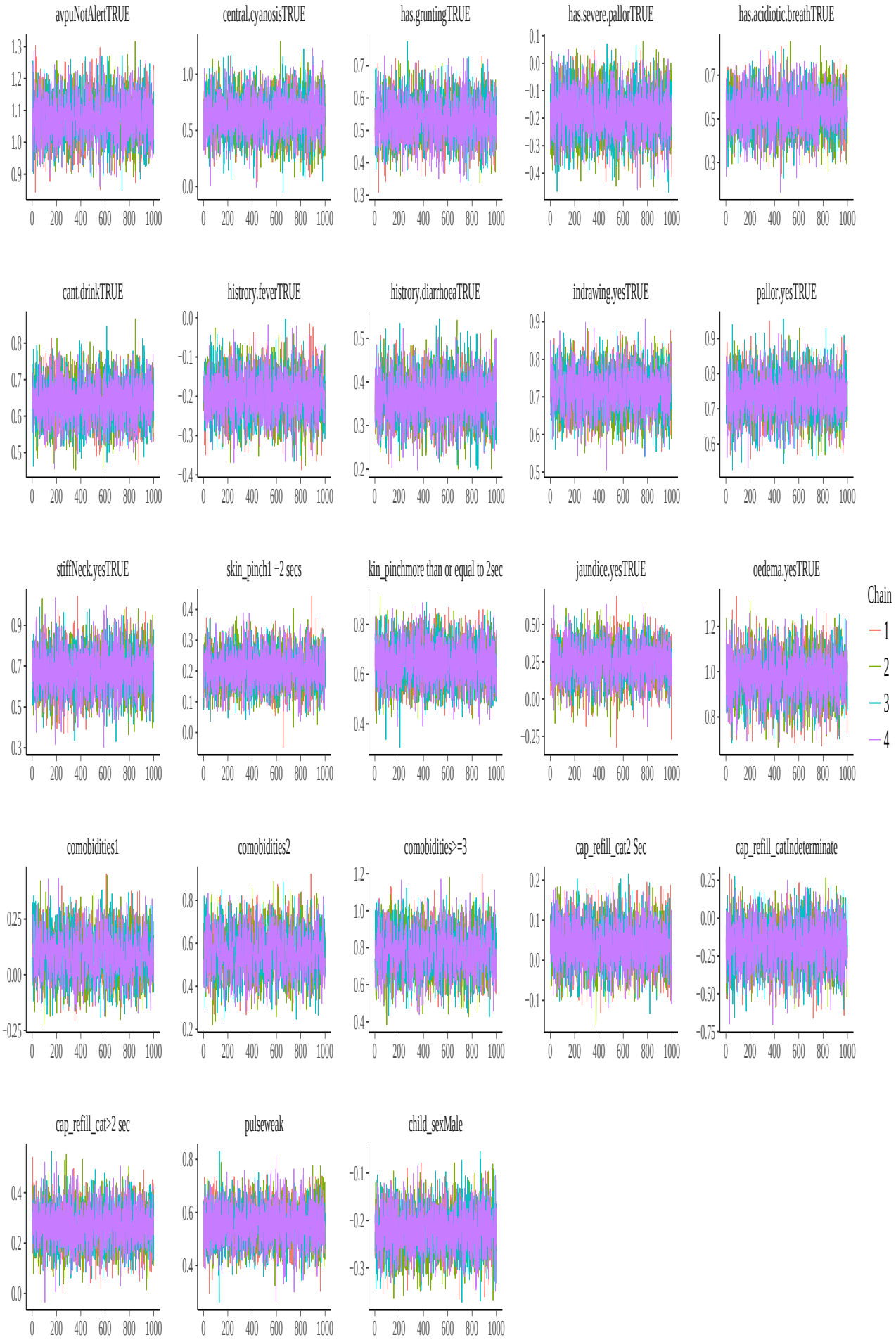

Figure 4.6: Convergence of prognostic model

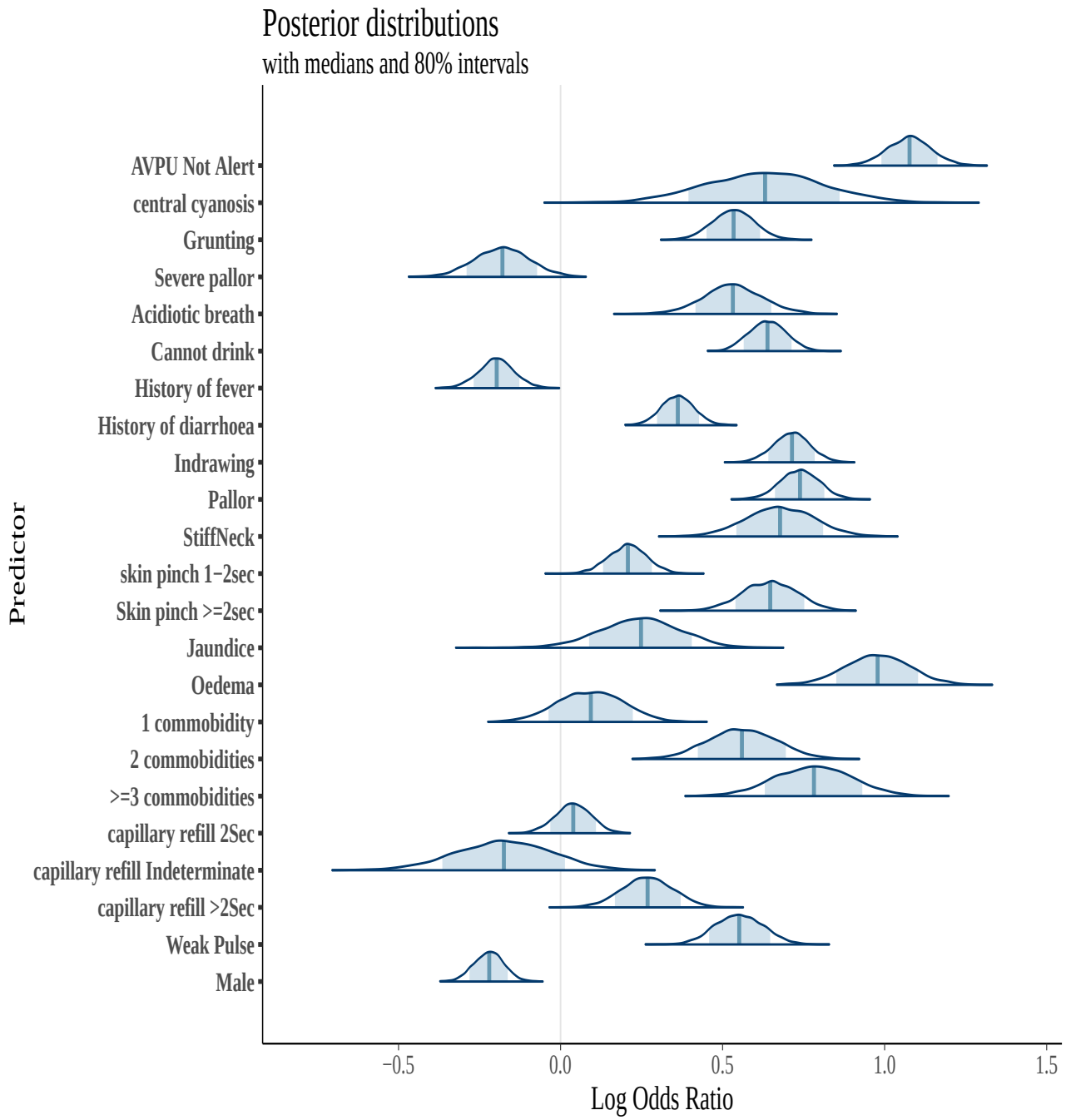

Figure 4.7: Posterior distribution of the model predictors

4.3 Posterior predictive density

4.3.1 Distributions of test statistics

Following the recommendation by (Gelman et al., 2014) for assessing the posterior predictive distribution for the test statistic such as mean and standard deviation, we simulated data from the posterior distribution and compared the distribution of mean in the generated simulation with the mean in the observed outcome as shown in Figure 4.8. The value of test statistic T which is computed from the observed data is shown as a vertical dark line $T(y)$. The plot also shows the vast majority of the simulated data sets under the model having mean and standard deviation which approximately equal to the observed data. Hence we concluded that our model made realistic predictions.

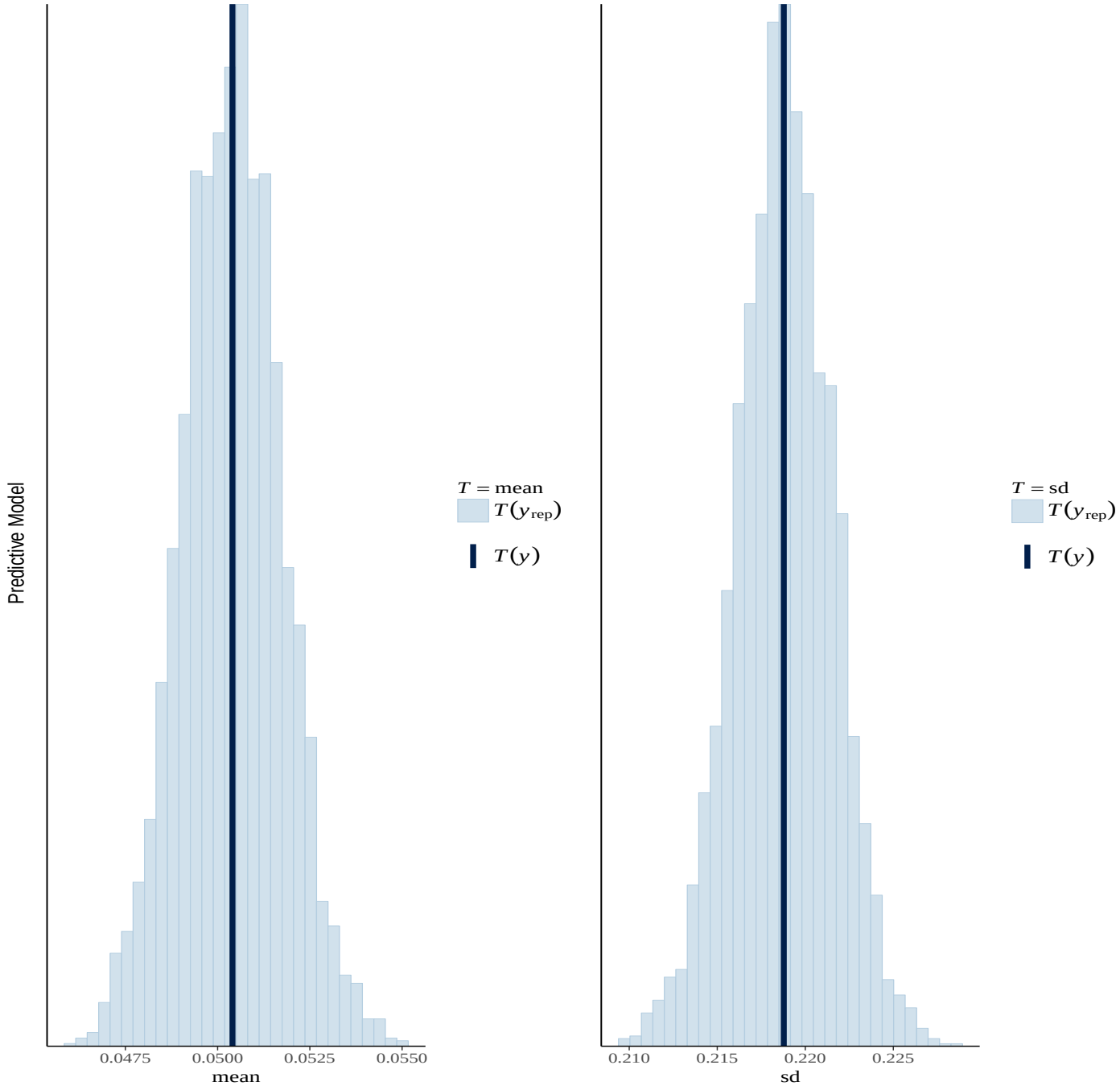

Figure 4.8: Posterior predictive distribution for the test statistic

Chapter 5

Discussion

5.1 Discussion

5.1.1 Principal findings in relation to the literature

Using a robust Bayesian approach we derived a new prognostic risk tool to be used in Kenya as well as other low-in-come countries(LIC). Bayesian approach allowed us to make more accurate predictions with the data that was available. To the best of our knowledge, this is the largest such study that utilizes routine data from multiple hospitals over a span of 3 years to have ever been undertaken in a predefined paediatrics population in LIC. Using the posterior predictive distribution of the derived model, we simulated data and comparative key statistics such as mean, standard deviation, minimum, a maximum predicted values as recommended by (Gelman et al., 2014). Results are as shown in Figure 4.8 which made us conclude that the derived model has a better calibration in out-of sample data. Furthermore, we compared the empirical distribution of the observed data to the distributions of simulated data from the posterior predictive distribution. The smoothed kernel density estimate seemed identical as shown in Figure 5.1 an indicator of a well calibrated model. However, since this validation was done on the simulated data of the derived model, we are aware that it potentially has some limitation. Consequently, an external validation has to be performed to assess transportability of this model before it is deployed in clinical practice.

The final model was to some extent complex given that it included more variables than recently derived models such as (George et al., 2015) and (Berkley et al., 2003). But this com-

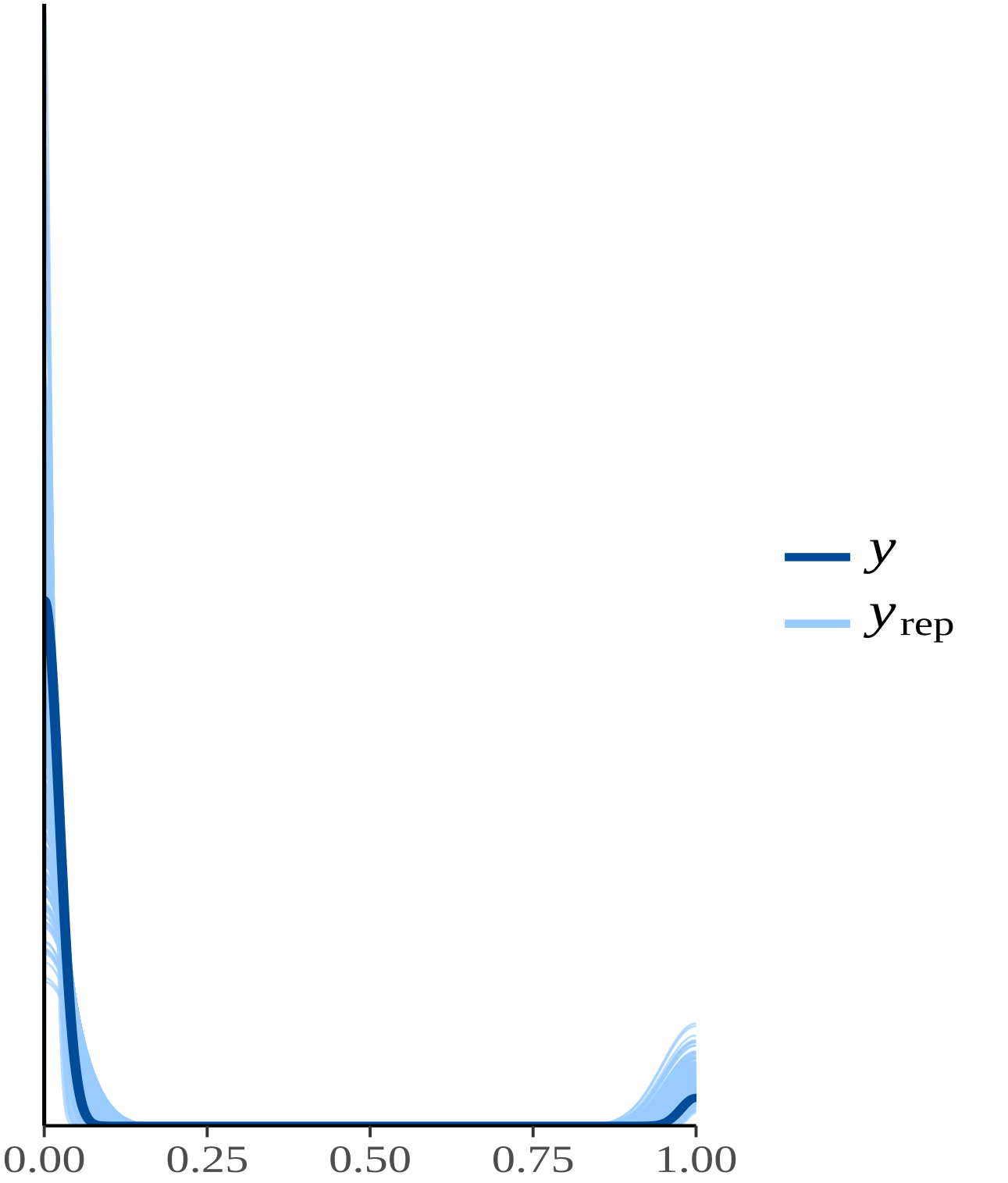

Figure 5.1: Comparison of the distribution of the observed data vs simulated data from the posterior predictive density

plexity was necessary considering the fact that the chosen variables are routinely collected at the bedside which do not require laboratory confirmation. Furthermore, the number of variables used was justified with the number of events in the current study hence events-per-variable rule as described by (Ogundimu et al., 2016) was not violated. Furthermore, no statistical procedures were used such as step-wise methods as applied in the cited studies to arrive at the final variables, rather the selection was based on the knowledge of the experts in paediatrics. This approach has been endorsed by (Wynants et al., 2016) and (Ogundimu et al., 2016) who have discouraged the automatic selection of features based on statistical significance.

For the derived models to be useful in the out-of-sample data, the derivation dataset has to be of better quality both in dependent variable and the prognostic factors. Although variables such as vital signs (*respiratory rate, pulse rate*) were amongst the listed predictors of mortality, we chose not to include them in the final model because we suspected a presence of digit-preference by the observer; theses indicators are manually counted and sometimes approximated to a nearest number as shown in the Appendix 1. Thus their quality is not guaranteed. Exclusion of these variables, therefore meant that our estimates will not be limited to only patients where these vital signs are manually counted.

Current prognostic models tend to over-predict or under-predict during external validations. Most of them are not well calibrated for the paediatric population and they make a huge assumption that patients are alike! An assumption which is not entirely correct. As demonstrated in this study, there exists subgroups that underlie the population under study. These subgroups arise as a result of latent variables or some other random mechanisms at play, which the traditional statistical methods fail to take into account hence the utility of mixture model in this current study to identify observations that consists of subgroups as shown in Figure 4.2. We have demonstrated that a model that included subgroups had the better predictive ability than a model without. To achieve that we used leave-one-out cross-validation methods for Bayesian models as described by (Vehtari et al., 2016).

The derived model included hospital identity as one of the random effect. This decision improved the fit and was justified by the fact that hospitals under study were located in different geographical locations as shown in Figure 3.1 which serve different populations(rural/urban), different malaria endemicity(high/low) and also have different mortal-

ity rates as shown in Figure 4.1.

Crepitations or indrawing as is commonly referred to was the single most respiratory distress that the strongest predictor of mortality. These findings support the WHO recommendation of using danger signs at admission as a proxy indicator of probable pneumonia diagnosis. This observation coincides with the findings by (George et al., 2015).

5.1.2 Future work

Laboratory test results were not considered as one of the potential prognostic factors. This is because laboratory results might not be easily available at admission especially for emergence cases. Consequently, we consider this as a leeway for the future studies to include laboratory tests as prognostic factors for the inpatient population. Furthermore, time-to-event analysis should be considered as an extension of this work to come up with hazards of deaths using prognostic factors identified in this study.

5.1.3 Strengths and limitations

One crucial limitation is that, despite vitals signs *(temperature, respiratory rate, pulse rate)* being shown as a potential prognostic factor in studies elsewhere, such as (George et al., 2015), this current study was persuaded to drop them from the list of useful risk factors based on the strongest evidence of digit heaping. Another limitation was the use of diagnostic episodes as one of our predictors, these episodes were solely based on the clinician's impression (not a laboratory test) so the possibility of false positives could not be ruled out.

One of the main strengths is that, this study was based on a large and a representative pediatric population with data collected over time from different hospitals. Data abstraction from a patient file had a series of rigorous data validation checks as detailed by (Ayieko et al., 2015). This means that the data used in this study was of high quality. The use of multiple imputation increased the power of this study since we were able to utilize all available data without discarding any as is usually the case in complete-case analysis.

Chapter 6

Conclusion and Recommendations

6.1 Conclusion

Standard customary statistical methods have been successful to a larger extent in explanatory analysis. But the same is not true when applied in developing predictive models, particularly when using larger complex medical datasets. Given that patient are not alike, a statistical methodology that clusters patients into homogeneous subpopulations should be used to account for the inherent variability in the medical patients. Computational methodology such as mixture models should be used to identify inherent subpopulations that underlie the population of medical patients under study. Once each observation has been linked to its corresponding subpopulation, a mixed effect prognostic model can be built where subpopulations will serve as a random-effect (clustering variable). The resulting coefficients will approximate a true data generating model.

Despite this current study being based on simple routine clinical signs, it has shown a promise of better predictive ability. Thus, our results need to be externally validated to assess the transportability of the model, particularly in a non-malaria regions because we suspected the presence of malaria parasitemia in a larger proportion of derivation dataset. We believe that this prognostic model could be invaluable in assessing the clinical care.

Bibliography

Aitkin, M., Anderson, D., and Hinde, J. (1981). Statistical modelling of data on teaching styles. *Journal of the Royal Statistical Society. Series A (General)*, pages 419–461.

Alkema, L., Chou, D., Hogan, D., Zhang, S., Moller, A.-B., Gemmill, A., Fat, D. M., Boerma, T., Temmerman, M., Mathers, C., et al. (2016). Global, regional, and national levels and trends in maternal mortality between 1990 and 2015, with scenario-based projections to 2030: a systematic analysis by the un maternal mortality estimation inter-agency group. *The Lancet*, 387(10017):462–474.

Ayieko, P., Ogero, M., Makone, B., Julius, T., Mbevi, G., Nyachiro, W., Nyamai, R., Were, F., Githanga, D., Irimu, G., et al. (2015). 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*, pages archdischild–2015.

Berkley, J., Ross, A., Mwangi, I., Osier, F., Mohammed, M., Shebbe, M., Lowe, B., Marsh, K., and Newton, C. (2003). Prognostic indicators of early and late death in children admitted to district hospital in kenya: cohort study. *Bmj*, 326(7385):361.

Betancourt, M. (2013). Generalizing the no-u-turn sampler to riemannian manifolds. arxiv, 1304.1920. *URL http://arxiv. org/abs/1304.1920*.

Bleeker, S., Moll, H., Steyerberg, E., Donders, A., Derksen-Lubsen, G., Grobbee, D., and Moons, K. (2003). External validation is necessary in prediction research:: A clinical example. *Journal of clinical epidemiology*, 56(9):826–832.

Collins, G. S., de Groot, J. A., Dutton, S., Omar, O., Shanyinde, M., Tajar, A., Voysey, M., Wharton, R., Yu, L.-M., Moons, K. G., et al. (2014). External validation of multivariable

prediction models: a systematic review of methodological conduct and reporting. *BMC medical research methodology*, 14(1):40.

Collins, G. S., Mallett, S., Omar, O., and Yu, L.-M. (2011). Developing risk prediction models for type 2 diabetes: a systematic review of methodology and reporting. *BMC medicine*, 9(1):103.

Collins, G. S., Ogundimu, E. O., Cook, J. A., Manach, Y. L., and Altman, D. G. (2016). Quantifying the impact of different approaches for handling continuous predictors on the performance of a prognostic model. *Statistics in medicine*, 35(23):4124–4135.

Cruz, J. A. and Wishart, D. S. (2006). Applications of machine learning in cancer prediction and prognosis. *Cancer informatics*, 2.

Duncan, H., Hutchison, J., and Parshuram, C. S. (2006). The pediatric early warning system score: a severity of illness score to predict urgent medical need in hospitalized children. *Journal of critical care*, 21(3):271–278.

Egdell, P., Finlay, L., and Pedley, D. (2008). The paws score: validation of an early warning scoring system for the initial assessment of children in the emergency department. *Emergency Medicine Journal*, 25(11):745–749.

Gathara, D., Malla, L., Ayieko, P., Karuri, S., Nyamai, R., Irimu, G., Hensbroek, M. B., Allen, E., and English, M. (2017). Variation in and risk factors for paediatric inpatient all-cause mortality in a low income setting: data from an emerging clinical information network. *BMC pediatrics*, 17(1):99.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2014). *Bayesian data analysis*, volume 2. Chapman & Hall/CRC Boca Raton, FL, USA.

Gelman, A., Lee, D., and Guo, J. (2015). Stan a probabilistic programming language for bayesian inference and optimization. *Journal of Educational and Behavioral Statistics*, page 1076998615606113.

Gelman, A. and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical science*, pages 457–472.

George, E. C., Walker, A. S., Kiguli, S., Olupot-Olupot, P., Opoka, R. O., Engoru, C., Akech, S. O., Nyeko, R., Mtove, G., Reyburn, H., et al. (2015). Predicting mortality in sick african children: the feast paediatric emergency triage (pet) score. *BMC medicine*, 13(1):174.

Harris, P. A., Taylor, R., Thielke, R., Payne, J., Gonzalez, N., and Conde, J. G. (2009). Research electronic data capture (redcap)—a metadata-driven methodology and workflow process for providing translational research informatics support. *Journal of biomedical informatics*, 42(2):377–381.

Helbok, R., Kendjo, E., Issifou, S., Lackner, P., Newton, C. R., Kombila, M., Agbenyega, T., Bojang, K., Dietz, K., Schmutzhard, E., et al. (2009). The lambarene organ dysfunction score (lods) is a simple clinical predictor of fatal malaria in african children. *Journal of Infectious Diseases*, 200(12):1834–1841.

Hoeting, J. A., Madigan, D., Raftery, A. E., and Volinsky, C. T. (1999). Bayesian model averaging: a tutorial. *Statistical science*, pages 382–401.

Hunsberger, S., Albert, P. S., and London, W. B. (2009). A finite mixture survival model to characterize risk groups of neuroblastoma. *Statistics in medicine*, 28(8):1301–1314.

Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V., and Fotiadis, D. I. (2015). Machine learning applications in cancer prognosis and prediction. *Computational and structural biotechnology journal*, 13:8–17.

Kucukelbir, A., Ranganath, R., Gelman, A., and Blei, D. (2015). Automatic variational inference in stan. In *Advances in neural information processing systems*, pages 568–576.

Kusmakar, S., Muthuganapathy, R., Yan, B., O'Brien, T. J., and Palaniswami, M. (2016). Gaussian mixture model for the identification of psychogenic non-epileptic seizures using a wearable accelerometer sensor. In *Engineering in Medicine and Biology Society (EMBC), 2016 IEEE 38th Annual International Conference of the*, pages 1006–1009. IEEE.

Le, T. Q. and Bukkapatnam, S. T. (2016). Nonlinear dynamics forecasting of obstructive sleep apnea onsets. *PloS one*, 11(11):e0164406.

Lu, P., Xia, J., Li, Z., Xiong, J., Yang, J., Zhou, S., Wang, L., Chen, M., and Wang, C. (2016). A vessel segmentation method for multi-modality angiographic images based on multi-scale filtering and statistical models. *Biomedical engineering online*, 15(1):120.

Mallett, S., Royston, P., Waters, R., Dutton, S., and Altman, D. G. (2010). Reporting performance of prognostic models in cancer: a review. *BMC medicine*, 8(1):21.

McLachlan, G. and Peel, D. (1998). Robust cluster analysis via mixtures of multivariate t-distributions. *Advances in pattern recognition*, pages 658–666.

Mikolajczyk, R. T., DiSilvesto, A., and Zhang, J. (2008). Evaluation of logistic regression reporting in current obstetrics and gynecology literature. *Obstetrics & Gynecology*, 111(2, Part 1):413–419.

Muro, S., Takemasa, I., Oba, S., Matoba, R., Ueno, N., Maruyama, C., Yamashita, R., Sekimoto, M., Yamamoto, H., Nakamori, S., et al. (2003). Identification of expressed genes linked to malignancy of human colorectal carcinoma by parametric clustering of quantitative expression data. *Genome biology*, 4(3):R21.

Oermann, E. K., Rubinsteyn, A., Ding, D., Mascitelli, J., Starke, R. M., Bederson, J. B., Kano, H., Lunsford, L. D., Sheehan, J. P., Hammerbacher, J., et al. (2016). Using a machine learning approach to predict outcomes after radiosurgery for cerebral arteriovenous malformations. *Scientific reports*, 6:21161.

Ogundimu, E. O., Altman, D. G., and Collins, G. S. (2016). Adequate sample size for developing prediction models is not simply related to events per variable. *Journal of clinical epidemiology*, 76:175–182.

Pollack, M. M., Ruttimann, U. E., and Getson, P. R. (1988). Pediatric risk of mortality (prism) score. *Critical care medicine*, 16(11):1110–1116.

Rajaratnam, J. K., Marcus, J. R., Flaxman, A. D., Wang, H., Levin-Rector, A., Dwyer, L., Costa, M., Lopez, A. D., and Murray, C. J. (2010). Neonatal, postneonatal, childhood, and under-5 mortality for 187 countries, 1970–2010: a systematic analysis of progress towards millennium development goal 4. *The Lancet*, 375(9730):1988–2008.

Schlattmann, P. (2009). *Medical applications of finite mixture models*. Springer Science & Business Media.

Schlattmann, P., Dietz, E., and BOeHNING, D. (1996). Covariate adjusted mixture models and disease mapping with the program dismapwin. *Statistics in Medicine*, 15(7-9):919–929.

Shann, F., Pearson, G., Slater, A., and Wilkinson, K. (1997). Paediatric index of mortality (pim): a mortality prediction model for children in intensive care. *Intensive care medicine*, 23(2):201–207.

Steyerberg, E. W. and Vergouwe, Y. (2014). Towards better clinical prediction models: seven steps for development and an abcd for validation. *European heart journal*, page ehu207.

Vehtari, A., Gelman, A., and Gabry, J. (2016). Practical bayesian model evaluation using leave-one-out cross-validation and waic. *Statistics and Computing*, pages 1–20.

Wakaba, M., Mbindyo, P., Ochieng, J., Kiriinya, R., Todd, J., Waudo, A., Noor, A., Rakuom, C., Rogers, M., and English, M. (2014). The public sector nursing workforce in kenya: a county-level analysis. *Human resources for health*, 12(1):6.

WHO (2005). *Handbook IMCI: integrated management of childhood illness*. World Health Organization.

Wynants, L., Collins, G., and Van Calster, B. (2016). Key steps and common pitfalls in developing and validating risk models. *BJOG: An International Journal of Obstetrics & Gynaecology*.

6.2 Appendix

.0.1

Mixed effect model used to extract linear predictors. Hospitals used as random effect

```
functions {
}
data {
int<lower=1> N;  // total number of observations
int Y[N];  // response variable
int<lower=1> K;  // number of population-level effects
matrix[N, K] X;  // population-level design matrix
// data for group-level effects of ID 1
int<lower=1> J_1[N];
int<lower=1> N_1;
int<lower=1> M_1;
vector[N] Z_1_1;
}
transformed data {

}
parameters {
vector[K] b;  // population-level effects
vector<lower=0>[M_1] sd_1;  // group-level standard deviations
vector[N_1] z_1[M_1];  // unscaled group-level effects
}
transformed parameters {
// group-level effects
vector[N_1] r_1_1;
vector[N] eta;
```

```
r_1_1 = sd_1[1] * (z_1[1]);
eta = X * b ;
for (n in 1:N) {
eta[n] = eta[n] + (r_1_1[J_1[n]]) * Z_1_1[n]; //Linear predictor
}
}
model {
// prior specifications
for( k in 2:K){
b[k]~ cauchy(0, 2.5); //Gelman 2008 (Coefficients)
}
b[1] ~ cauchy(0, 10); //Gelman 2008 (Intercept)

sd_1 ~ student_t(3, 0, 10);
z_1[1] ~ normal(0, 1); //Random effects

// likelihood contribution
Y ~ bernoulli_logit(eta);
}
generated quantities {
}
```

.0.2

Mixture model used to identify patients in subgroups

```
functions {
}
data {
int<lower=1> k_groups; // number of mixture components
int<lower=1> N; // number of data points
//real y[N]; // observations
```

```
vector[N] y;
}
parameters {
simplex[k_groups] theta; // mixing weights
ordered[k_groups] mu; //ordered for identifiability
real<lower=0> sigma[k_groups]; // scales of mixture components
}
model {
//Contribution from each component
real contributions[k_groups]; // temp for log component densities
//priors
sigma ~ cauchy(0, 2.5);
mu ~ normal(0, 10);
theta ~ dirichlet(rep_vector(5,k_groups));
//likelihood
for (n in 1:N) {
for (k in 1:k_groups) {
contributions[k] = log(theta[k]) + normal_lpdf(y[n] | mu[k], sigma[k]);
}
target += log_sum_exp(contributions);
}
}
//Abstraction of probabilities of each observation belonging to a given mixture componen
generated quantities {
matrix[N, k_groups] p;
for (n in 1:N){
vector[k_groups] ps;
for (K in 1:k_groups){
ps[K] = log(theta[K]) + normal_lpdf(y[n] | mu[K], sigma);
}
```

```
p[n, ] = transpose(softmax(ps));
}
}
```

.0.3

Mixed effect model with hospitals & subpopulations as random effects

```
functions {
}
data {
int<lower=1> N;  // total number of observations
int Y[N];  // response variable
int<lower=1> K;  // number of population-level effects
matrix[N, K] X;  // population-level design matrix
// data for group-level effects of ID 1
int<lower=1> J_1[N];
int<lower=1> N_1;
int<lower=1> M_1;
vector[N] Z_1_1;
// data for group-level effects of ID 2
int<lower=1> J_2[N];
int<lower=1> N_2;
int<lower=1> M_2;
vector[N] Z_2_1;
int prior_only;  // should the likelihood be ignored?
}
transformed data {
int Kc;
matrix[N, K - 1] Xc;  // centered version of X
vector[K - 1] means_X;  // column means of X before centering
```

```
Kc = K - 1;  // the intercept is removed from the design matrix
for (i in 2:K) {
means_X[i - 1] = mean(X[, i]);
Xc[, i - 1] = X[, i] - means_X[i - 1];
}
}
parameters {
vector[Kc] b;  // population-level effects
real temp_Intercept;  // temporary intercept
vector<lower=0>[M_1] sd_1;  // group-level standard deviations
vector[N_1] z_1[M_1];  // unscaled group-level effects
vector<lower=0>[M_2] sd_2;  // group-level standard deviations
vector[N_2] z_2[M_2];  // unscaled group-level effects
}
transformed parameters {
// group-level effects
vector[N_1] r_1_1;
// group-level effects
vector[N_2] r_2_1;
r_1_1 = sd_1[1] * (z_1[1]);
r_2_1 = sd_2[1] * (z_2[1]);
}
model {
vector[N] mu;
mu = Xc * b + temp_Intercept;
for (n in 1:N) {
mu[n] = mu[n] + (r_1_1[J_1[n]]) * Z_1_1[n] + (r_2_1[J_2[n]]) * Z_2_1[n];
}
// prior specifications
sd_1 ~ student_t(3, 0, 10);
```

```
z_1[1] ~ normal(0, 1);
sd_2 ~ student_t(3, 0, 10);
z_2[1] ~ normal(0, 1);
// likelihood contribution
if (!prior_only) {
Y ~ bernoulli_logit(mu);
}
}
generated quantities {
real b_Intercept;  // population-level intercept
b_Intercept = temp_Intercept - dot_product(means_X, b);
}
```

.0.4

End-Digit preference of the vital signs readings at admission

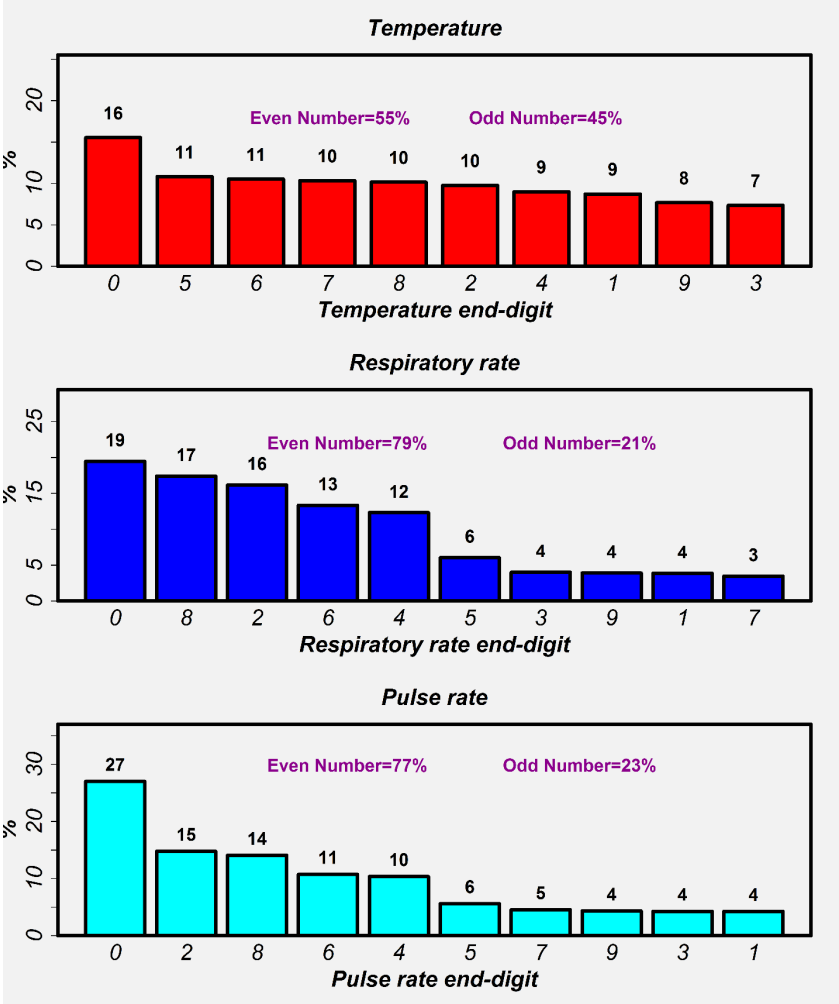

Figure 1: End-Digit preference of the vital signs readings at admission