COLLEGE OF NATURAL AND COMPUTATIONAL SCIENCES

DEPARTMENT OF STATISTICS

Modeling of respiratory rates of preterm infants admitted to neonatal intensive care units in Ethiopia

By

Fetene Kebede

A thesis submitted to the Department of Statistics in the partial fulfillment of the requirements for the Degree of Master of Science in Statistics (Biostatistics)

June, 2020

Addis Ababa

Modeling of respiratory rates of preterm infants admitted to neonatal intensive care units in Ethiopia

MSc Thesis

By:

Fetene Kebede

Advisor: Bedilu Alamirie (PhD**)**

June, 2020

Addis Ababa

Statement of Author

I declare that this thesis is a result of my work and all sources of materials used for writing it have been duly acknowledged. I have submitted this thesis to Addis Ababa University in partial fulfillment of the requirements of Degree of Master of Science in Statistics (Biostatistics) complies with the regulations of the university and meets the accepted standards with respect to originality and quality. I declare that I have not so far submitted this thesis to any other institution anywhere for the award of any academic degree, diploma or certificate.

Fetene Kebede

Date: _____

Signature: _____

June 2020

Addis Ababa, Ethiopia

As the members of the board of examiners of MSc thesis open defense examination, we certify that we have read and evaluated the thesis and examined the candidate. Hence, we recommend that the thesis be accepted as it fulfills the requirements for the Degree of Master of Science in Statistics (Biostatistics).

Chairman _____ **Signature** _____ **Date** _____

Advisor _____ **Signature** _____ **Date** _____

Examiner _____ **Signature** _____ **Date** _____

Examiner _____ **Signature** _____ **Date** _____

TABLE OF CONTENTS

Acknowledgements

First of all I would like to thank the almighty God for the gift of health and strength throughout my life and in the process of carrying out this study and for seeing me through all. I wish to acknowledge the support of all who contributed in various ways to make this study a success. My deep gratitude goes to my advisor Dr. Bedilu Alamirie, for his valuable suggestions and comments to the successful realization of this study.

Secondly, I gratefully acknowledge the Study Illness of Preterm project staff members for providing different suggestion and all the relevant data used in this study. Thirdly, my heartfelt thanks go to my parents for their support in the period of my study and thesis work. Finally, I would like to extend my gratitude to all who support me directly and indirectly during successful realization of this study.

LIST OF TABLES

LIST OF FIGURES

ACRONYMS

ALR	Alternating logistic regression
AIC	Akaike's information criterion
BPD	Bronchopulmonary dysplasia
CI	Confidence interval
GA	Gestational Age
GEE	Generalized estimating equation
GLM	Generalized linear model
GLMM	Generalized linear mixed model
ML	Maximum likelihood
NICU	Neonatal intensive Care Unit
OR	Odds-ratio
SIP	Study illness of preterm
WHO	World Health Organization

Addis Ababa University

College of Natural Science and Computational Sciences

Department of Statistics

Degree of Master of Science in Statistics (Biostatistics)

Modeling of Respiratory Rates of Preterm Infants Admitted to Neonatal Intensive Care Unit in Ethiopia

By Fetene Kebede

June, 2020

ABSTRACT

Respiratory rate problem is the major reason of morbidity and mortality in infancy and childhood which has a long-term impact on health outcomes in adult life. The main objective of the study was identifying risk factors associated with preterm infant respiratory rate. Secondary data was taken from 2586 infants admitted to the neonatal intensive care unit of five hospitals in three regions of Ethiopia from July 1, 2016 to May 31, 2018. The result of the study showed that, the magnitude of preterm infants who were born with abnormal respiratory rate was 1270(49.1%). According to generalized estimating equation modeling we arrived at the finding results. Controlling all other variables in the model, female preterm child were 13% less likely to have abnormal respiratory rate than male. (aOR= 0.869, 95% CI: 0.758, 0.996). Similarly, preterm infants who have extremely preterm were 85% more likely to have abnormal respiratory rate than moderate preterm (aOR= 1.855, 95% CI: 1.385, 2.481) and preterm child who have very preterm were 60% more likely to have abnormal respiratory rate than moderate preterm (aOR= 1.603, 95% CI: 1.220, 2.106). Those preterm infants who have higher birth weight were less likely to have abnormal respiratory rate; those preterm infants who have 1500-2500 grams were 27% less likely to have abnormal respiratory rate than below 1500grams (aOR= 0.733, 95% CI: 0.604, 0.889). According to generalize linear mixed model the finding results given the same random intercept $b_{j,}$ female preterm child were 14% less likely to have abnormal respiratory rate than male. Similarly, very preterm and extremely very preterm were 66% and 97% more likely to have abnormal respiratory rate than moderate preterm respectively. Child born with birth weight between 1500-2500 were 18% less likely to have abnormal respiratory rate than birth weight less than 1500grams and child born mother who from diabetic disease were 99% more likely abnormal respiratory rate than mothers who are not diabetic disease. In conclusion, infant sex, gestational age, and birth weight were statistically significant factors of respiratory rate.

CHAPTER ONE

1. Introduction

1.1 Background

Heart rate and respiratory rate are key vital signs used to assess the physiological status of children in many clinical settings (National Collaborating Centre for Women's and Children's Health, 2013). They are used as initial measurements in acutely unwell children, as well as in those undergoing more intensive monitoring in high dependency or intensive care settings. During cardiopulmonary resuscitation, heart rate and respiratory rate are critical values used to determine responses to lifesaving interventions. Heart rate and respiratory rate remain an integral part of the standard clinical assessment of children presenting with acute illnesses (National Collaborating Centre for Women's and Children's Health, 2013)

Respiratory rate is the number of breaths per minute or more formally, the number of movements indicative of inspiration and expiration per unit time. In practice, the respiratory rate is usually determined by counting the number of times the chest rises or falls per minute. The aim of measuring respiratory rate is to determine whether the respirations are normal or abnormal (fast, slow). The significance of this sign was recognized that rapid respirations were a "sign of pain or inflammation" (Hooker et al., 1989)

Preterm birth is defined by WHO as all births before 37 completed weeks of gestation or fewer than 259 days since the first day of a woman's last menstrual period (WHO, 1997). Globally from 139.95 million live births of the total born child in 2014, an estimated preterm birth rate is 14.84% babies were born preterm (WHO, 2014). Over 80% of preterm births occur in sub-Saharan Africa and South Asia (WHO, 2014). Ethiopia is among the top 15 countries that contribute to two-thirds of the world's preterm babies with an estimated preterm birth rate of 14.1% (Blencowe et al., 2010) and the top 10 countries that contribute to two-thirds of the world's preterm babies with an estimated preterm birth rate of 12% (WHO, 2014).

Preterm birth can be further sub-divided based on gestational age: extremely pret**e**rm (<28 weeks), very preterm (28 - <32 weeks) and moderate preterm (32 - <37 completed weeks of gestation). The systematic review, by (Been et al., 2014) investigated respiratory symptoms resulting after preterm birth. Preterm-born

children had an increased risk of wheezing disorders with an unadjusted OR of 1.71 (95% CI 1.57–1.87) compared with term-born children. Very preterm born children <32 weeks' gestation had an unadjusted OR of 3.00 (95% CI 2.61–3.44). The respiratory symptoms increased with increasing prematurity (Edwards et al., 2016).

Respiratory problem has multiple factors for Preterm infant whose solutions for factors from an array of discoveries that will not come through a single discovery but it addressing multiple biological, clinical, and social behavioral risk factors. Respiratory problem preterm causal factors linked to include medical conditions of the mother or fetus related, genetic influences, condition or environmental exposure, prior infertility treatments, behavioral factors as well as iatrogenic prematurity (Hibbard et al., 2010).

In Ethiopia the Study illness of preterm project has the task to determine the most common causes of illness and mortality in preterm infants admitted to hospitals in Ethiopia based on a standardized diagnostic protocol. The major consequences of prematurity are respiratory problems; however the respiratory squeal heterogeneity and varying definitions have made overall a systematic assessment difficult. The following variables are considered as possible risk factors for respiratory problem in the neonatal period: maternal age, chronic maternal diseases, renal disorders, diabetes mellitus, ulcerative colitis, and systemic lupus erythematous and chronic hypertension, mode of delivery, antenatal steroid treatment, infant sex and gestational age in completed weeks (Muhe et al., 2019).

According to Study illness of preterm project in Ethiopia between July 1, 2016 and May 31, 2018, 4,919 preterm infants were enrolled in the study and 3,852 were admitted to neonatal intensive care units. By 28 days of post-natal age, 1,109 (29%) of those admitted to the neonatal intensive care unit died. The main cause of death in the 1109 infants was established as respiratory problem (45%). To treat these conditions further research is required to determine the most common causes of illness (respiratory problem) to prevent and treat the major causes of preterm respiratory problem (Muhe et al., 2019).

Studies have been done regarding this not so much in Ethiopia, because of in Ethiopia due to limited information on data of respiratory rate. It is important to determine the risk factor of respiratory rate in various

hospitals in the country in order to come up with feasible intervention strategies by creating awareness about factors associated with respiratory rate of preterm infant to minimize the problem.

1.2 Statement of the problem

Preterm birth is one of the major causes of infants' death as compared to term infants which is not an acute disease and experience more difficulty with feeding, temperature instability and respiratory case. Respiratory problem of preterm children is one of the critical problems in Ethiopia which causes many infants short- term and long-term health which can result higher mortality and morbidity. As far as the knowledge of the current investigators is concerned there is no enough investigation on factors which are significantly affecting the infant preterm respiratory system. The information gap is boldly observed on the area of factors affecting respiratory rates of preterm infants while the investigators of the current study are reviewing different literature.

Due to lack of information, the number of infant mortality and morbidity are increasing from time to time. The health professionals need full information and evidence to follow advice and treat their patients to significantly decrease the magnitude of infant mortality and morbidity due to preterm respiratory problem. There is not enough scientific evidence on infant preterm respiratory problem as a result conducting scientifically sound full research on modeling of respiratory rates of preterm infants is very important.

So, to fill the observed gap the current study focuses on modeling of respiratory rates of preterm infants and identify risk factors associated with preterm infant respiratory rate using secondary data from the Study Illness of Preterm project to contribute a valuable information and conclusion for health professionals.

1.3 Objective of the study

The main objective of the study is to identify risk factors associated with preterm infant respiratory rate by applying suitable statistical models.

1.4 Significance of the study

Respiratory problem of Preterm infant is influenced by multiple factors whose solutions will not come through a single discovery but rather from an array of discoveries addressing multiple biological, clinical, and social behavioral risk factors. There is need for longitudinal studies for long term and short term to report the respiratory outcomes of preterm-born subjects from infancy through to later life. So this is a longitudinal study and the results of this study will also be useful in creating awareness about factors associated with respiratory rate of preterm infant and reducing preterm infant mortality.

Specifically:

- The results of the study may contribute to the understanding determinant factors that affecting the respiratory rate of preterm infant in Ethiopia.
- To give emphases on the factors that has strong association with respiratory rate of preterm infant so that decision directives act on it accordingly.
- The study can be used as a stepping-stone for further studies

CHAPTER TWO

2. Literature review

Preterm born is an important indicator of its vulnerability to the risk of respiratory problem childhood illnesses and the chances of survival. Children whose preterm born birth is considered to have a higher risk of early childhood death due to respiratory problem than children whose term born birth (National Collaborating Centre for Women's and Children's Health, 2013).

Respiratory rate is the number of breaths per minute or, more formally, the number of movements indicative of inspiration and expiration per unit time. Breathing is a frequent feature of respiration in the newborn and is especially common in the premature infant. Although prematurely newborns are closely monitored because complications and disease are expected, low-risk infants born at term usually go through a minimum of standardized observations as long as they appear good. The major consequences of prematurity are respiratory problems, and bronchopulmonary dysplasia (BPD) is the most common respiratory consequence of premature birth and contributes to significant short- and long-term morbidity, mortality, and resource utilization (Northway et al., 1967).

Wen et al. (2019) conducted a study in Taiwan with main objective to determine the relationship between maternal preeclampsia and neonatal respiratory distress syndrome (RDS) from the database of the premature baby foundation of Taiwan retrospective cohort study with data from very-low-birth-weight (VLBW) infants born in 1997–2014.The method used for this study is multivariate logistic regression to analyze the association between maternal preeclampsia and RDS risk. The study with multivariate logistic regression analysis result indicated that, gestational age (GA), birth weight, infant sex, SGA, and antenatal steroid use were considered as potential confounders, whereas preeclampsia was not significantly associated with RDS (adjusted odds ratio (aOR), 1.12; 95% CI, 0.98–1.29). GA and birth weight were associated inversely with severe RDS, SGA and antenatal use of two or more doses of steroids protected against the development of severe RDS [aOR (95% CI), 0.8 (0.69–0.93) and 0.57 (0.53–0.62), respectively]. Male infants were more likely to develop RDS as compared to females and severe RDS [aOR (95%CI), 1.18 (1.06–1.32) and 1.24 (1.14–1.35), respectively].

Zanardo et al. (2009) conducted a study in Padua University, Padua Italy neonatal intensive care unit from January 2001 to December 2006, having the objective to assess whether histological chorioamnionitis (HCA), in the setting of preterm premature rupture of membranes (PPROM), affects infant respiratory outcome. Prospective histological study was conducted on the 287 NICU admitted preterm infants, 68/287 (23.6%) the results of the study was identified HCA, 16/68 (23.5%) with a coexisting fetal inflammatory response, and 74/287 (25.7%) with PPROM. Infant respiratory outcome was significant related with a greater frequency of vaginal delivery (P< 0.0001), lower gestational age (P< 0.0001) and lower birth weight (P< 0.01). HCA had insignificant effect on fetal lung maturation; however, it was a significant factor for CLD (RR; 95% CI 2.08; 1.30–3.33). HCA and similarly fetal inflammatory response were also significant risk factors for PPROM (RR; 95% CI 2.07; 1.42– 3.03 and 2.64; 1.71–4.09 respectively). Conversely, HCA in the setting of PPROM failed to reveal any RDS protection or subtype CLD risk. Multivariate analysis demonstrated significant effects of presence of maternal HCA (P= 0.04), gestational age (P< 0.0001) and interaction HCA-gestational age (P= 0.04) on CLD development, regardless of the presence of fetal HCA or fetal HCA-gestational age interaction, PPROM or PPROM-gestational age interaction.

Altman et al. (2013) conducted a study based on data about all infants born in 2004–2008 in Sweden, with objective to establish risk factors for acute respiratory rates diseases in moderately preterm infants and the method used in this study was logistic regression and the result of study showed that, maternal, obstetric and neonatal as the risk factors for the two most common diagnoses, transient tachypnea of the newborn (TTN) and respiratory distress syndrome (RDS). The summary result of the study included 4679 moderately preterm [gestational age (GA): 30 to 34 weeks], 15 036 late preterm infants (GA 35 to 36 weeks) and 451 479 term infants (GA: 37 to 41 weeks). In moderately preterm infants, risk factors for TTN in multivariable analyses were multiparity, caesarean section before and after onset of labour, male sex, Apgar score 4–6 at 5 min and lower GA. Risk factors for RDS were multiparty, caesarean section before and after onset of labour, if sex of child male, Apgar score<7 at 5 min and lower GA. Preterm rupture of membranes, antenatal corticosteroid treatment and being small for gestational age reduced the risk of RDS.

Anna et al. (2018) conducted a retrospective study in Poland to investigate the cause of respiratory distress syndrome (RDS) in neonates from singleton pregnancies from January 2011 to December 2014 aimed with

preterm premature rupture of membranes (pPROM) between 24+0 and 36+6 weeks by using regression analysis for various factors. Totally in 175 singleton pregnancies include with pPROM, 95 cases of RDS (54, 29%) was diagnosed. latency period of PROM, gestational age at birth, Umbilical Artery Pulsatility Index (UA PI), Middle Cerebral Artery Pulsatility Index (MCA PI), fetal distress, antenatal steroids use, delivery type, pregnancy hypertension disease, gestational glucose intolerance or diabetes, neonatal laboratory parameters, gender, weight, Apgar score, and other neonatal complications in all cases the following information was collected. to investigate the effect of variables on RDS the method used is logistic regression analysis. The results of logistic regression analysis showed that: female gender (OR=0.52; 95%CI:0.28-0.97), antenatal steroids use (OR=0.46; 95%CI:0.34-0.64), abnormal UA PI and MCA PI (OR=2.96; 95%CI:1.43-6.12) (OR=2.05; 95%CI:1.07-3.95), fetal distress (OR=2.33; 95%CI:1.16-4.71), maternal HGB (OR=0.69; 95%CI:0.5-0.96), and neonatal RBC, HGB (OR=0.32; 95%CI:0.19-0.55) (OR=0.75; 95%CI:0.65-0.88).The main RDS risk factors in premature neonates are gender, abnormal fetoplacental circulation, and fetal distress. The laboratory parameters such as lower RBC and HGB count are observed in infants with RDS.

Debelew et al. (2014) conducted a study on the prospective follow-up with the objective to identify determinants factors neonatal mortality and causes of mortality for neonatal in Jimma Zone, Southwest Ethiopia on 3463 neonates from September 2012 to December 2013. The method used for this study was mixed-effects multilevel logistic regression model. The findings indicate that the status of neonatal mortality rate was about 35.5 (95%CI: 28.3, 42.6) per 1000 live births. Though significant variation was existed between clusters in relation to neonatal mortality, cluster-level variables were found to have insignificant effect on neonatal mortality. Individual-level variables such as birth order, frequency of antenatal care use, delivery place, gestation age at birth, premature rupture of membrane, complication during labor, twin births, size of neonate at birth and neonatal care practice were identified as determinants of neonatal mortality. Birth asphyxia (47.5%), neonatal infections (34.3%) and prematurity (11.1%) were the three leading causes of neonatal mortality accounting for 93%.

Alemayehu et.al (2016/2017) conducted a study in southern Tigray, Ethiopia, with the objectives to assess the risk factors of acute respiratory infections among under-five children attending public hospitals. An institutional based unmatched case control study design was employed. Binary logistic regression was

employed to assess statistical association via odds ratio. Significance of statistical association was assured or tested using 95% confidence interval and *P*-value (0.05). Bivariate and multivariate analyses were employed to examine the relationship or statistical association between the outcome variable and selected independent variables. The study showed that malnutrition (AOR = 2.89; 95%CI: 1.584–8.951; $p = 0.039$), cow dung use (AOR =2.21; 95%CI: 1.121–9.373; $p = 0.014$), presence of smoker in the family (AOR = 0.638; 95% CI: 0.046–0.980; $p = 0.042$) and maternal literacy (AOR = 3.098; 95%CI: 1.387–18.729; $p = 0.021$) were significant predictors of acute respiratory infection among under-five children.

Respiratory problem is a serious health problem caused by factors that are potentially modifiable. Therefore, it is very important to determine the risk factor of respiratory problem in various hospitals in the country in order to come up with feasible intervention strategies to minimize the problem. In this study the corresponding outcome variables are categorical and count responses, responses variables will be correlated within individuals in the same clusters. There may be also having cluster variations within and between hospitals on respiratory problem. To handle such types of data, the most flexible and appropriate models should be applied.

CHAPTER THREE

3. Data and methods

3.1 Source of data

Data was collected retrospectively on a total of 2586 infants admitted to the Neonatal Intensive Care Units in five hospitals in three regions of Ethiopia from July 1, 2016 to May 31, 2018 we used in this study. They were obtained from Gandhi memorial hospital (271), Tikur anbessa hospital (683) and St. Paul hospital (705) in Addis Ababa, Gondar university hospital (678) in the north and Jimma university hospital (249) in south-west of Ethiopia. All preterm infants who had at least two visits follow up and measured respiratory rates are eligible for the study. Neonates admitted to neonatal intensive care unit who have incomplete records were not included.

3.1.1 Study population

The population includes premature infants admitted to neonatal intensive care unit in one of the above hospitals with a gestational age of less than 37 completed weeks.

3.1.2 Variables considered in the study

The respiratory rate of preterm infant was reported as normal and abnormal. These two groups form the categories of outcome variable. The binary outcome variable Y generated for the occurrence of respiratory rate of preterm infant as:

$$\boldsymbol{Y_{ij}} = \begin{cases} \mathbf{1}, & \text{Normal , if respiratory rate is between 40 and 60 per minute} \\ \mathbf{0}, & \text{Abnormal} \end{cases}$$

The collected data consist of Y_{ij} for the observed respiratory status where Y_{ij} is the outcome for the i^{th} child (i =1, 2,…, 2586) at measurement time j (j =1, 2, …, 28) days.

Table: 3.1 Coding and explanation of explanatory variable

Attributes	**Description**		**Categories**	
Sex	Sex of child	0=Female	1=Male	
Birth weight	Weight of child at birth	0=Less than 1500	1= Between 1500-2500	2= Greater than 2500
Gestation age	Gestation age of child in weeks	1= Extremely preterm (<28 weeks)	2= Very preterm (28 to 32 weeks)	3= Moderate preterm (>32&<37 weeks)
Age	Age of mother at time of delivery	1=15-24 years	2=25-34 years	3=35-49 years
Antibiotics	Maternal medication prior to delivery	0=No	1=Yes	
Antenatal Visits	Antenatal visit during pregnancy	1=No Visit	2=Yes	
Hypertation disease	Pregnancy hypertation disease	1=Yes	2=No	
Pregnancy	Number of children	0=Single	1=Twins	2= Triplet
Delivery mode	Methods of childbirth	1= Spontaneous vaginal delivery 4= Assisted breech	2= Forceps 5= C-Section	3= vacuum
Cardiac disease		1= Yes	2= No	
Diabetes mellitus		1=Yes	2= No	

3.2 Data analysis

For this study, some extension of generalized linear models such as marginal models and cluster specific models will be applied. Data was entered in to SAS software for analysis.

3.3. Generalized linear models (GLM)

Generalized linear models (GLMs) extend ordinary regression models to encompass non normal response distributions and modeling functions of the mean (Agresti, 2002). Three components that specify a generalized linear model are random component, which identifies the response variable Y and its probability distribution; a systematic component specifies explanatory variables were used in a linear predictor function; and a link function specifies the function of expected value of the response variable that the model equates to the systematic component. In general, GLM is a linear model for a transformed mean of a response variable that has distribution in the natural exponential family.

The Exponential Family: A random variable Y follows a distribution that belongs to the exponential family, if the density function is of the form

$$\boldsymbol{f(y/\theta,\phi) = exp\{a(\phi)^{-1}[y\theta - \psi(\theta)] + c(y,\phi)\}} \qquad (\mathbf{3.1})$$

For a specific set of unknown parameters θ and ϕ, and for known functions $\psi(\cdot)$ and $c(\cdot,\cdot)$. The parameter θ is called the canonical parameter and represents the location, while ϕ is called the dispersion parameter and represents the scale parameter. An important property of the GLM is the functional relation between mean and variance.

3.3 Marginal models

Observations are usually taken from the same unit in clustered data, and thus this information indicates a cluster of correlated observations. Proper analysis of clustered data is required in statistical modeling the association between the dependent or response variable and the given set of independent covariates. Marginal models are among the most statistical models widely used to model clustered or repeated data. The marginal model primary objective is to analyze the given factors in the study on the binary response variable of interest of the population-averaged effects. This means that the marginal expectations and the covariates are directly related (Molenberghs & Verbeke, 2005). The marginal models fitted in this study using generalized estimating equations (GEE).

3.3.1 Generalized estimating equations (GEE)

For binary data, a GEE approach is an extension of GLMs. It provides a semi-parametric approach to longitudinal analysis of categorical response; it can be also used for continuous measurements. GEE approach is used to account for the correlation between responses of interest for subjects from the same cluster (Diggle et al., 1994). GEE is non-likelihood method that uses correlation to capture the association within clusters or subjects in terms of marginal correlations (Molenberghs & Verbeke, 2005). For clustered as well as repeated measured data, (Liang & Zeger, 1986) proposed GEE which require only the correct specification of the univariate marginal distributions provided one is willing to adopt "working" assumptions about the correlation structure. The "working" assumptions as proposed by Liang and Zeger; included independence, unstructured, exchangeable and autoregressive. Independence and exchangeable working assumptions can be used in virtually all applications, whether longitudinal, clustered, multivariate, or otherwise correlated. Autoregressive and unstructured correlation structures are less relevant for clustered data, studies with unequally spaced measurements or sequences with differing lengths (Molenberghs and Verbeke, 2005).

Let $y_j = \left(y_{j1}, \dots \dots y_{jnj}\right)'$ be the response values of observations from j^{th} cluster,

for $j = 1,2, \dots m$ follows a binomial distribution i.e $y_j = Bin(n_j, \pi_j)$ that belongs to the exponential family with the density function of the form (3.1). Then, to model the relation between the response and covariates, one can use a regression model similar to the generalized linear models given by:

$$g(\pi_j) = logit(\pi_j) = X_j'\beta \qquad (3.2)$$

where $g(\pi_j)$=logit link function, $X_j = (n_j \times P)$dimensional vector of covariates,

$\beta = (1 \times P)$ dimensional vector of unknown fixed regression parameter to be estimated and $E(Y_j) = \pi_j$ is expected values of the j^{th} response variable from a hospital.

Parameter estimation for GEE

As previously expressed GEE is not likelihood approach, rather it is quasi-likelihood based and estimates β by solving estimating equations which consist of the working covariance matrix V_j. The score equation that used to estimate while accounting for the correlation structure for the marginal regression parameters is:

$$S(\beta) = \sum_{j=1}^{m} \frac{\partial \pi_j}{\partial \beta'} \left[A_j^{1/2} R_j A_j^{1/2}\right]^{-1} (Y_j - \pi_j) = 0 \qquad (3.3)$$

where R_j working correlation matrix and the covariance matrix of are Y_j is decomposed in to

$$A_j^{1/2} R_j A_j^{1/2}$$

with A_j the matrix with the marginal variances on the main diagonal and zeros elsewhere and Y_j is multivariate vector of asymptotically normal response variables with mean vector π_j

i.e. $Y_j \cong N(X_j\beta, V_j)$ is a consistent estimator of $\hat{\beta}$, GEE model approach is an advantage, even when structure of the working correlation matrix, R_j is not correctly defined or misspecified. However, working correlation structure for severe misspecification may seriously affect that the efficiency of the GEE model estimators (Molenberghs & Verbeke, 2005).

3.3.2 Marginal model building

Model selection is basic and an important issue in almost any activities of practical data analysis. One of the data analysis strategy is model selection. Model selection is used to search of best models. This means selecting the best subset of the covariates from the available covariates in the data.

3.3.3 Variable selection technique

The selection of significant variables in the GEE model, start by fitting model that contains all covariates for model building strategy. Considerations about exchangeable, independence, unstructured and autoregressive working correlation have been made. The backward elimination procedure was used to select important factors related to respiratory rate. The strategy is called backward because of working backward direction starting from full model to a reduced model. This indicates that covariates with insignificant p-values are removed step by step. The variables with the p-value greater than 0.05 will be removed sequentially and new model will be refitted with the remaining covariates. Finally, two models will be compared using model comparison techniques. It turned out that the model with significant covariates is found to be the most parsimonious model (Zorn, 2001).

3.3.4 Model comparison technique

Quasi-Information Criterion (QIC): In a condition, when the likelihood function cannot be fully specified, such as in the GEE case, the Akaike's Information Criterion (AIC) cannot be directly applied to select either the optimal set of explanatory variables or correlation matrix. As an alternative, one can use the modified Akaike's Information Criterion called Quasi Information Criteria (QIC), which is based on the quasi-likelihood function (Pan, 2001). QIC is derived from the AIC and conceptually similar. The quasi-likelihood function takes the following form (McCullagh &Nelder, 1989)

$$Q(\pi) = \int_{y}^{\pi} \frac{y-t}{\phi_v(t)} dt$$

where $\pi = E(y), v(y) = \phi_v(\pi)\ and\ \phi$ the dispersion parameter. An equation for the QIC is $IC = -2Q(\hat{\pi}, I) + 2trace\left[(\Omega_I^{-1}\hat{V}_R)\right]$, where I represent the independent correlation structure (diagonal matrix) and R is the specified working correlation structure.

The p-dimensional matrices Ω_I^{-1} and $\hat{V}_R$ are variance estimators of the regression coefficients under the correlation structure I and R respectively. The QIC value will be computed based on the quasi-likelihood estimate $\hat{\pi}$ and will be used to select the candidate explanatory variables. The model with the smallest QIC value for all correlation structures will be considered as the best candidate model.

3.3.5 Model checking technique

Preisser and Qaqish (1996) models for the correlated data was fitted further diagnostics generalize regression by generalized estimating equations (GEEs), where the influence of measured observations entire clusters correlated. For marginal models the proposed diagnostic measures were similar to those that exist for generalized linear models: DFBETAC, Cluster Cooks '*D*, Cluster leverage and Cluster DFFIT. The diagnostic purpose of each measure is similar as well. DFBETAC is a measure of the influence that any cluster has on each $\hat{\beta}$ (Belsley et al., 1980); Cluster Cooks' *D* is a measure of the influence of any cluster on the overall fit of the model (Cook, 1982); Cluster leverage is a measure of how extreme cluster is with respect to the predictors (Belsley *et al*., 1980). Cluster DFFIT represents the studentized Cook distance type statistic to measure the influence of deleting cluster on the overall model fit. DFBETAC, Cluster Cooks 'D and Cluster DFFIT are referred to as deletion diagnostics because the magnitude of each is related to changes in the fit of the model after a particular cluster is removed compared to the fit of the model on the full data. Let n_i be the number of responses for cluster i, and $N = \sum_{i=1}^{k} n_i$ the total number of observations. A_i is n_i x n_i diagonal matrix Let B NxN diagonal matrix and let B_i the n_i x n_i diagonal matrix corresponding to cluster i. Let $Q_i = X_i(X'X)^{-1}X_i'$ where X_i is the n_i x p design matrix corresponding to cluster i. The adjusted residual vector is defined as $E = B(Y - \hat{\pi})$ and $E_i = B_i(Y_i - \hat{\pi}_i)$ the estimated residual for the i^{th} cluster.

CLEVERAGE

The leverage of cluster i is contained in the matrix $H_i = Q_i$ and is summarized by the trace of H_i, where H_i is the hat matrix of cluster i.

$$\text{CLEVERAGE}_i = tr(H_i)$$

The leverage value greater than one for the i^{th} cluster indicates that cluster is influential (Belsley et al., 1980).

DFBETAC

The effect of deleting cluster i on the estimated parameter vector is given by the following one-step approximation for $\hat{\beta} - \hat{\beta}_{[i]}$:

$$\text{DFBETAC}_i = (X'X)^{-1}X_i'(I - Q_i)^{-1}E_i$$

If $\mathrm{DFBETAC}_i$ is less than unity, this implies no specific impact of cluster on the coefficient of a particular predictor variable, while $\mathrm{DFBETAC}_i$ of i^{th}cluster greater than 1.0, implies the cluster is an outlier (Cook and Weisberg, 1982).

DFBETACS

The cluster deletion statistic DFBETAC can be standardized by dividing the components of DFBETAC by its standard error.

CLUSTERCOOKSD

Let $\mathrm{DCLS}_{\boldsymbol{i}}$ be the cluster-level Cook's D for cluster i, which can be calculated as

$\mathrm{DCLS}_i = E_i'(I - Q_i)^{-1})Q_i(I - Q_i)^{-1})E_i/p\hat{\phi}$ Where p is the number of predictors in the model and $\hat{\phi}$ is dispersion parameter. The suggested cut off values for i^{th} cluster to be influential is, if DCLS_i is greater than "one" (Preisser and Qaqish, 1996).

CLUSTERDFFIT

Let MCLS_i be the cluster-level DFFIT for cluster i which can be calculated as

$\mathrm{MCLS}_i = E_i'(I - Q_i)^{-1}H_iE_i/p\hat{\phi}$ The suggested cut off values for i^{th} cluster to be influential is, if MCLS_i is greater than "one" (Preisser and Qaqish, 1996).

PEARSON RESIDUAL

Another model diagnostic tool for marginal model is Pearson residual. Raw residuals and Pearson residuals are available for models fit with generalized estimating equations (GEEs). The raw residual is defined as

$$r_i = y_i - n\hat{\pi}_i$$

Where y_i is the i^{th} response and $n\hat{\pi}_i$ corresponding predicted mean.

The Pearson residual is defined by the difference between observed and fitted values and divides by an estimate of the standard deviation of the observed value. Observations with a Pearson residual exceeding three in absolute value may shows lack of fit (Davison and Snell, 1991). Pearson residual is given by:

$$r_{pi} = \frac{y_i - n\hat{\pi}}{[var(y_i)]^{1/2}}$$

3.4 Generalized linear mixed model (GLMM)

Generalized linear models (GLM) are part of subject specific models which extends ordinary regression by allowing non-normal responses and a link function of the mean. The generalized linear mixed model is a further extension that permits random effects as well as fixed effects in the linear predictor (Agresti, 2002).
When interest is in the marginal or population-averaged models to analyze the relationships of the covariates to the dependent variable for an entire population, marginal models as discussed in previous section are preferred. However, in most biomedical and biological data problems, interest often lies in understanding the response of individual patient characteristics and how this response is influenced by a given set of possible covariates (Myers et al.,2010). This proves even to be essential when individual interventions may be necessary. Cluster specific models are useful in such cases. Cluster specific models differ from the marginal models by inclusion of parameters that are specific to clusters or subjects within a population. Consequently, random effects will directly use in modeling the random variation in the dependent variable at different levels of the data (Petersen, 2004).

Let y_{ij} denote the response of i^{th} individual child from j^{th} hospital where $i = 1,2,\dots,n_j$ and y_j the n_j dimensional vector of all measurements available for hospital j. Let $f(b_j/D)$ be the density of the $N(0,D)$distribution for the random effect b_j. Assumed conditionally on q-dimensional random effects b_j to be drawn independently from $N(0,D)$, the outcomes y_{ij} of Y_j are independent with the density of the form

$$f_j(y_{ij}/b_j,\beta,\phi) = exp\{a(\phi)^{-1}[y_{ij}\theta_{ij} - \psi(\theta_{ij})]+c(y_{ij},\phi)\} \qquad (3.4)$$

Then the generalized linear mixed model (Molenberghs and Verbeke, 2005); with logit link is defined as

$$logit(\pi_{ij}) = X'_{ij}\beta + Z'_{ij}b_j\,, \qquad j = 1,2,\dots,m \qquad (3.5)$$

where $E(Y_{ij}/b_j) = \pi_{ij}$, is the mean response vector conditional on the random effects b_j , for child in hospital j and, X_{ij} and Z_{ij} are p-dimensional and q-dimensional vectors of known covariate values. The random effects b_j are assumed to follow a multivariate normal distribution with mean 0 and covariance matrix D.

3.4.1 Parameter estimation for GLMM

Random-effects models can be fitted by maximization of the marginal likelihood, obtained by integrating out the random effects. Such likelihood may involve high-dimensional integrals that cannot be evaluated analytically. The likelihood of the data expressed as a function of unknown parameters (Booth & Hobert, 1999).

$$L(\beta, D, \phi) = \prod_{j=1}^{m} f_j\left(Y_j/\beta, D, \phi\right) = \prod_{j=1}^{m} \int \prod_{j=1}^{n_j} f_{ij}\left(Y_{ij}/b_j, D, \phi\right) f\left(b_j/D\right) db_j \qquad (3.6)$$

It is the integral over the unobserved random effects of the joint distribution of the data and random effects. The problem in maximizing (3.6) is the presence of m integrals over the q-dimensional random effects b_j With Gaussian data, the integral has a closed form solution and relatively simple methods exist for maximizing the likelihood or restricted likelihood. With non-linear models, numerical techniques are needed. The Laplace method (Mohlenberg's & Verbeke, 2005) will be designed to approximate integrals of the form:

$$I = \int e^{Q(b)}\, db \qquad (3.7)$$

where $Q(b)$ are a known, unimodal, and bounded function of a q-dimensional variable b. Let $\hat{b}$ be the value of b for which Q is maximized. Then the second order Taylor expansion of $Q(b)$ is the form

$$Q(b) \approx Q\left(\hat{b}\right) + \frac{1}{2}\left(b - \hat{b}\right)' Q''\left(\hat{b}\right)\left(b - \hat{b}\right) \qquad (3.8)$$

where, $Q''\left(\hat{b}\right)$ is the matrix of second-order derivative of Q, evaluated at $\hat{b}$.Replacing $Q(b)$ in (3.7) by its approximation in (3.8) we obtain

$$I \approx (2\pi)^{q/2}\left|-Q\left(\hat{b}\right)\right|^{-1/2} e^{Q(\hat{b})}$$

Clearly, each integral (3.6) is proportional to an integral of the form (3.7) for functions $Q(b)$ given by

$$Q(b) = \phi^{-1} \sum_{i=1}^{n_j} \left[y_{ij} \left(x'_{ij}\beta + Z'_{ij}b \right) - \psi(x'_{ij}\beta + Z'_{ij}b) \right] - \frac{1}{2} b' D^{-1}$$

Note that the $\hat{b}$ of Q depends on the unknown parameters β, ϕ, and D, such that in each iteration of the numerical maximization of the likelihood will be recalculated conditionally on the current values for the estimates for these parameter.

3.4.2 Model building for GLMM

A different approach to account for clustering is by using random components such as random intercepts (Zeger & Karim, 1991). Under the GLMM, model building will begin by adoption of the marginal model covariates. Additionally, the model will also include the random effects, in this case, random intercepts to address the between and within-hospital heterogeneity. These are introduced in the generalized linear mixed model due to the fact that, the probability of having respiratory problem baby possibly varies for individuals within the same hospital as well as individuals in different hospital. Variable selection procedure for GLMM is similar with marginal model previously explained.

3.4.3 Model comparison in GLMM

The likelihood ratio test and Akaike's information criterion are used to select the best model based on the values of asymptotic estimations.

Likelihood Ratio Test: In order to decide on the better of the two random effects models, two models will be fitted, one with the two random intercepts (between and within Hospital variations) and another with one random intercept (within Hospital variation). One can use the approximate restricted maximum likelihood ratio test (LRT) to compare these two models (Myers et al., 2010).

Let $\boldsymbol{LR_{full} = -2 \log likelihood}$ value for full model and

$\boldsymbol{LR_{redu} = -2\, log\, likelihood}$ value for reduced model.

Then, the likelihood ratio test statistic, is given by $\boldsymbol{\lambda = LR_{full} - LR_{redu}}$

The asymptotic null distribution of the likelihood ratio test statistic $\boldsymbol{\lambda}$, is a chi-square distribution with degrees of freedom equal to the difference between the numbers of parameters in the two models.

Akaike's information criterion (AIC): AIC is a measure of goodness of fit of an estimated statistical model. It is not a test on the model in the sense of hypothesis testing; rather it is a tool for model selection. The AIC penalizes the likelihood by the number of covariance parameters in the model, therefore

$$\boldsymbol{AIC = -2\log(L) + 2P}$$

Where, L is the maximized value likelihood function for the estimated model and p is the number of parameters in the model. The model with the lowest AIC value is preferable.

3.4.4 Model checking technique

In GLMM, it is assumed that the random effects are normally distributed and uncorrelated with the error term. Normality of the random effects is assessed using normal plot of each random effect. Normal Q-Q plot of estimated random effects is an important method for checking the normality (Myers et al., 2010).

CHAPTER FOUR

4. Results

4.1. Summary of descriptive statistics

Before any statistical analysis, it is necessary to examine the overall picture of the data of preterm infant. This section provides a description of the data by using different statistical tools like frequency, percentage mean, median; modes finally the data are summarized in Table 4.1.

A total of 2586 infants are from five hospitals of three regions of Ethiopia (Addis Ababa, Oromia and Amhara) were selected for this study. Among these eligible preterm born infant, 1316(50.9%) child were born with normal respiratory rate whereas 1270(49.1%) were born with abnormal respiratory rate. The majority 1381(53.4%) of mothers were between 24-35 age interval and 2425(93.8%) mothers received antenatal care. Almost half 1203(46.5%) of infants were female. 2556(98.9%) and 2553(98.8%) of the mothers were free from cardiac disease and diabetes mellitus, respectively. Moreover, 713(27.5%) of mothers were hypertensive during pregnancy. About 1699(65.7%) and 836(32.3%) of pregnancy types were single and twin, respectively. The majority 1607(62.1%) of delivery mode was spontaneous vaginal delivery (SVD) and 996(38.5%) of mothers received at least one antibiotics during pregnancy. More than half 1554(60.1%) of the infants had birth weight between 1500-2500 grams and 1659(64.2%) were of more than37weeks gestational age. The prevalence of infants with abnormal respiratory system were 1270(49.1%).

Table 4.1: Summary of descriptive statistics for respiratory rate of preterm infant.

Factors	**Respiratory rate**		**Total (percentage)**
	Normal total (percentage)	**Abnormal total (percentage)**	
Maternal age			
Less than 24	468(35.6)	480(37.8)	948(36.7)
Between 24-35	722(54.9)	659(51.9)	1381(53.4)
Above 35	126(9.6)	131(10.3)	257(9.9)
Antenatal care received			
Yes	1237(51.0)	1188(49.0)	2425(93.8)
No	79(49.1)	82(50.9)	161(6.2)
Cardiac disease			
Yes	17(56.7)	13(43.3)	30(1.1)
No	1299(50.8)	1257(49.2)	2556(98.9)
Diabetes mellitus			
Yes	16(48.5)	17(51.5)	33(1.2)
No	1299(50.9)	1254(49.1)	2553(98.8)
Hypertensive disorders of pregnancy			
Yes	361(50.6)	352(49.4)	713(27.5)
No	955(51.0)	918(49.0)	1873(72.5)
Is pregnancy			
Single	871(66.2)	828(65.2)	1699(65.7)
Twins	421(32.0)	415(32.7)	836(32.3)
Triplet	24(1.8)	27(2.1)	51(2.0)
Delivery Mode			

SDV	825(62.7)	782(61.6)	1607(62.1)
Forceps	13(1.0)	11(0.9)	24(0.9)
Vacuum	3(0.2)	2(0.2)	5(0.2)
Assisted Breech	3(0.2)	5(0.4)	8(0.3)
C-Section	472(35.9)	470(37.0)	942(36.4)
Antibiotics			
yes	504(50.6)	492(49.4)	996(38.5)
No	812(51.1)	778(48.9)	1590(61.5)
Infant Sex			
Female	652(49.5)	551(43.4)	1203(46.5)
Male	664(50.5)	719(56.6)	1383(53.5)
Birth weight in grams			
Less than 1500	437(33.2)	503(39.6)	940(36.3)
Between 1500-2500	830(63.1)	728(57.3)	1554(60.1)
Greater than 2500	52(4.0)	39(3.1)	92(3.6)
Best gestational age in weeks			
Extremely preterm (<28 weeks)	16(1.2)	35(2.8)	51(2.0)
Very preterm (28 to 32 weeks)	355(27.0)	521(41.0)	876(33.9)
Moderate preterm (<37weeks)	945(71.8)	714(56.2)	1659(64.2)
Total	1316(50.9)	1270(49.1)	2586

4.2 Statistical analysis of marginal models

In this section, respiratory rate problem will be analyzed using generalized estimating equation models.

Analysis of generalized estimating equations (GEE)

The GEE starts by fitting all possible covariates. The backward elimination procedure was used to select significant factors related to respiratory rate.

The full model used was given:

$$\begin{aligned}\boldsymbol{logit}(\boldsymbol{\pi_{ij}}) = {} & \beta_0 + \beta_1 Sex_{female} + \beta_2 GA_{<28} + \beta_3 GA_{28-32} + \beta_4 Motherage_{24-35} + \beta_5 Motherage_{35+} \\ & + \beta_6 Birthweight_{1500-2500} + \beta_7 Birthweight_{2500+} + \beta_8 Modedelevery_{forceps} \\ & + \beta_9 Modedelevery_{vacuum} + \beta_{10} Modedelevery_{Ass.breech} + \beta_{11} Modedelevery_{C-section} \\ & + \beta_{12} priorHD_{yes} + \beta_{13} Antibiotics_{yes} + \beta_{14} pregnancy_{twin} + \beta_{15} pregnancy_{triplet} \\ & + \beta_{16} antenatalcare_{yes} + \beta_{17} CardiacD_{yes} \\ & + \beta_{18} DiabetsM_{yes} \qquad (4.1)\end{aligned}$$

The fitting process of the full model removed the covariate with the largest p-value (greater than 0.05) from the equation and the model was then refitted sequentially accordingly with the rest of the other covariates. Antenatal care received, mode of delivery, prior antibiotics, maternal age, prior hypertensive disorder, pregnancy number, cardiac disease and diabetic disease were excluded from the model step by step. It turned out that the model with infant sex, gestational age, birth weight was identified as the most parsimonious model.

Finally, comparison of empirical-based and model-based estimators of standard errors on the given working correlation assumptions (exchangeable, independence, unstructured and autoregressive) were obtained. The correlation structures that model-based standard errors and empirical-based standard errors are close to each other taken to be the best assumption correlation structure (Liang & Zeger, 1986).

Moreover, since Table 4.2 shows no differences among the correlations structure, using the exchangeable working correlation structure is the best assumption correlation structure. In addition, the empirical standard

errors for exchangeable correlation structures are somewhat smaller than their counterparts under the independence, unstructured and autoregressive assumptions.

Then, from Table 4.2 exchangeable working correlation assumption was found to be plausible since the two standard errors close to each other with correlation parameter (α =0.084). Then, the final model equation for generalized estimating equation model with statistically significant factors for respiratory rate is given as:

$$logit(\pi_{ij}) = \beta_0 + \beta_1 Sex_{female} + \beta_2 GestationalAge_{28-32} + \beta_3 GestationalAge_{less\ than\ 28} + \beta_4 Birthweight_{1500-2500} + \beta_5 Birthweight_{above\ 2500} \quad (4.2)$$

The final GEE model with parameter estimates and also with the corresponding empirically-based corrected standard errors including the p-values and adjusted odd ratio with 95%CI are presented in Table 4.2.

Table 4.2 Parameter estimates (empirical based corrected standard errors) for GEE based on model (4.2)

Factors		Estimate ($\hat{\beta}$)	Standard Error	P-value	aOR	95%CI aOR
Intercept		0.541	0.134	<.000	1.717	1.321, 2.234
Infant Sex	Male (ref)				1	
	Female	-0.140	0.070	0.044	0.869	0.758, 0.996
Gestational age in weeks	Moderate preterm (ref)				1	
	Extremely preterm	0.618	0.149	<.000	1.855	1.385, 2.481
	Very preterm	0.472	0.139	0.001	1.603	1.220, 2.106
Birth weight in	Less than 1500 (ref)				1	

grams	Between 1500-2500	-0.310	0.099	0.002	0.733	0.604, 0.889
	Greater than 2500	-0.319	0.183	0.081	0.727	0.507, 1.04

QIC=14780.537 α =0.084

Ref = reference category

The current study indicated that, infant sex, gestational age and birth weight significantly affect respiratory rate.

After controlling all other variables in the model female preterm infants were 13% less likely to have abnormal respiratory rate than male (aOR= 0.869, 95% CI: 0.758, 0.996). Similarly, those preterm infants who have extremely preterm or gestational age of below 28weeks (aOR= 1.855, 95% CI: 1.385, 2.481) were 85% more likely to have abnormal respiratory rate than those who have moderate preterm or above 32 weeks gestational age. Additionally, those infants who have very preterm or gestational age of 28-32weeks (aOR= 1.603, 95% CI: 1.220, 2.106) were 60% more likely to have abnormal respiratory rate than those who have moderate preterm or above 32 weeks gestational age.

Those preterm infants who have higher birth weight were less likely to have abnormal respiratory rate. In terms of category, those preterm infants who have 1500-2500 grams (aOR= 0.733, 95% CI: 0.604, 0.889) were 27% less likely to have abnormal respiratory rate than those who are below 1500grams Table 4.2.

4.3 Model diagnostic for marginal models

Plots of DFBETA, Cook's distance, leverage and cluster DFFIT value as a function of ordered cluster can then be used to see the pattern of all cases.

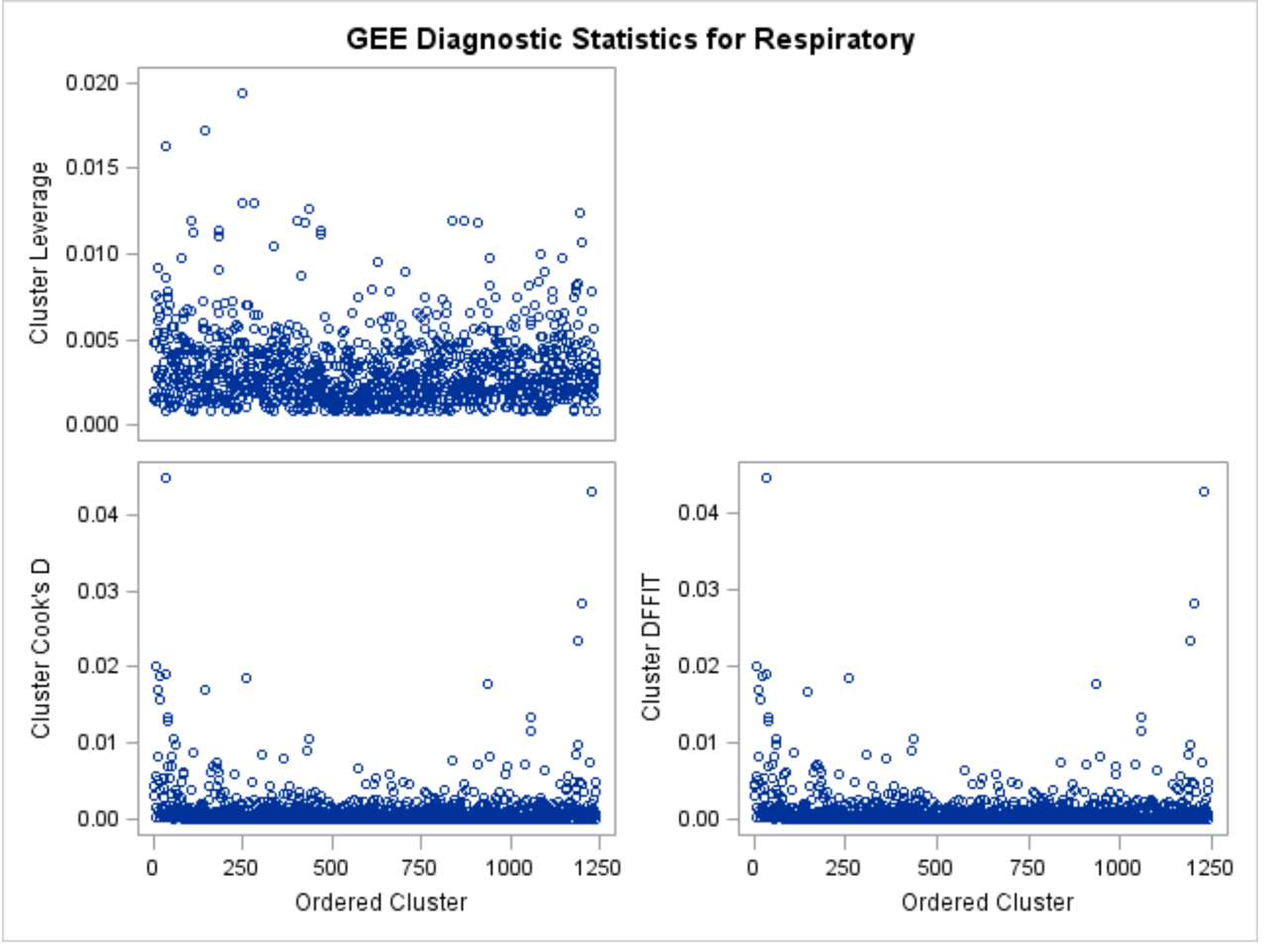

Figure 4.1: Plots of cluster leverage, cluster Cook's *D* and cluster DFFIT versus ordered cluster.

Figure 4.1 plots of leverage values versus the ordered cluster. It was observed that leverage values of the above plots are less than one. Therefore, there are no outliers.

Figure 4.1 also shows plot of Cook's *D* statistic versus the ordered cluster of all cluster. There are clusters a little far away from the others but these are not influential clusters since all Cook's *D* statistic are less than one.

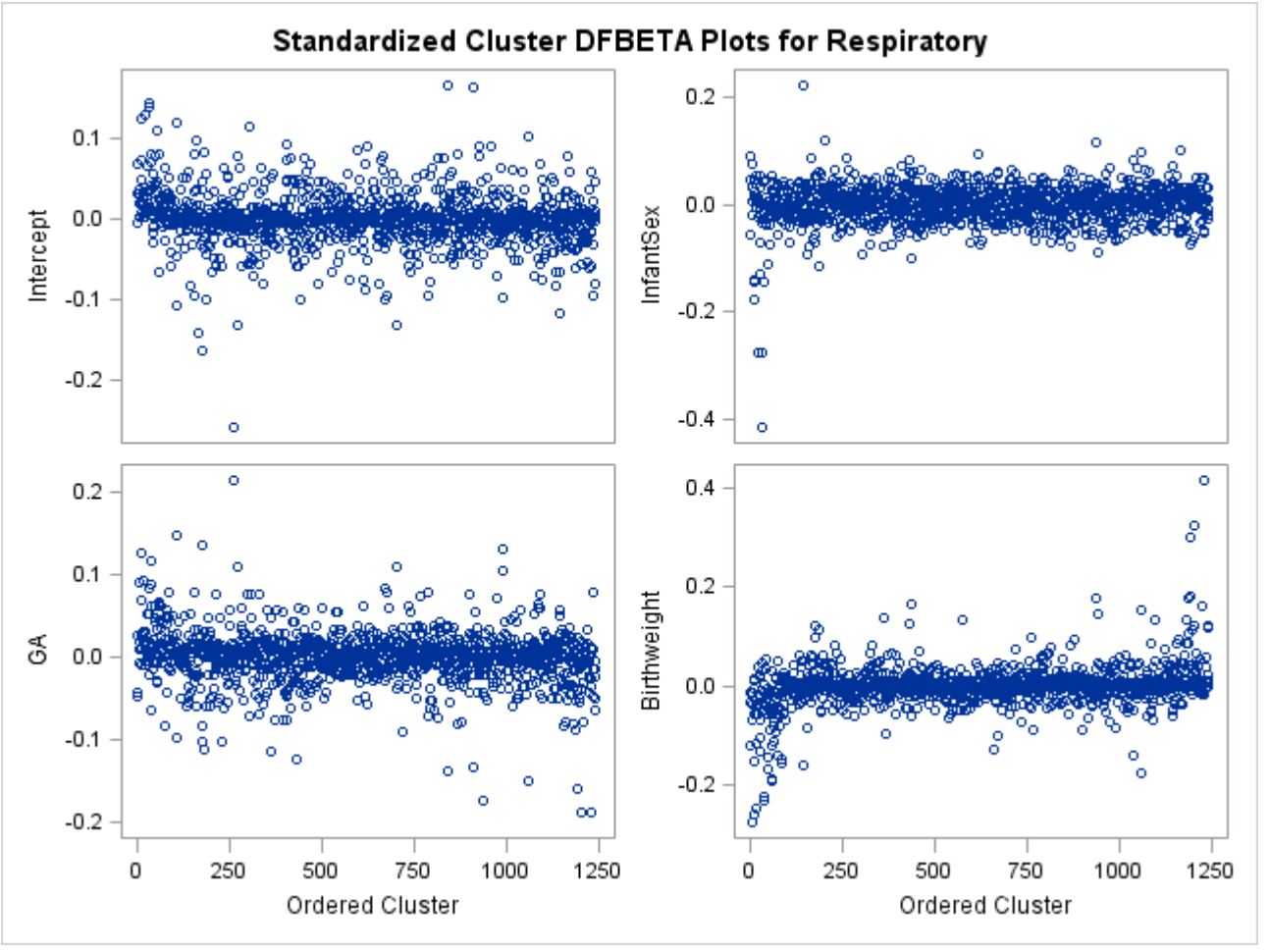

Figure 4.2: Plots of DFBETACS versus ordered cluster for all predictors in the fitted model.

Plots of DFBETACS of all explanatory variables versus order cluster are given in Figures 4.2 where it is shown that all the DFBETACS of all explanatory variables are less than one. This is an indication that there is no serious problem with the fitted model.

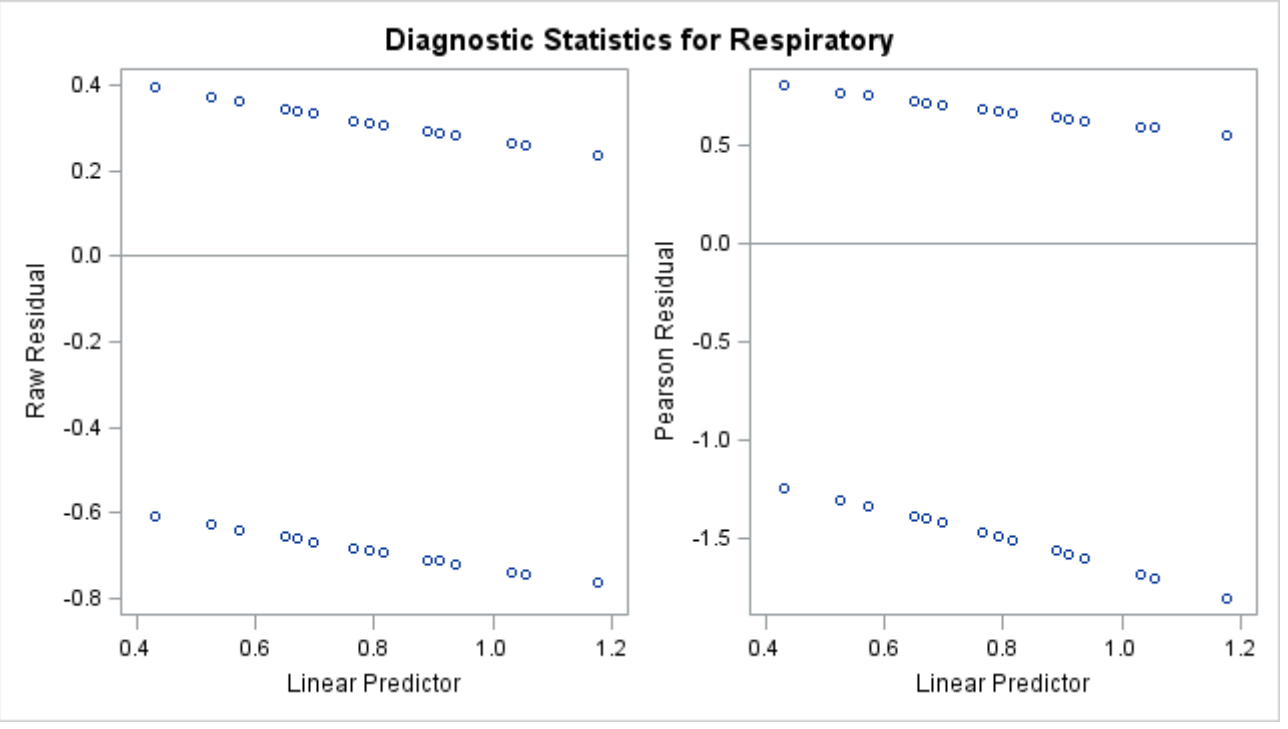

Figure 4.3: Plots of raw and Pearson residual versus linear predictors.

Figure 4.3 is the plot of raw residuals and Pearson residual versus linear predictors of all observations. There are few observations far away from others. However, the computed Pearson residuals do not influence the model that means all Pearson residuals are less than one.

4.4 Generalized linear mixed model (GLMM)

The GLMM, model fitting begins with adoption of the marginal model covariates. In addition to fixed effect considered in the marginal model, in the GLMM modeling framework, random effect terms were introduced to address the between-hospital and within-hospital variations. First, main effect covariates and two random intercepts models were fitted and as usual, non-significant covariates were removed sequentially starting from variables with highest p-value for fixed effect covariates. The saturated model for GLMM was fitted as follows where, b_j and b_{ij} two random intercepts

$$logit(\pi_{ij}) = \beta_0 + \beta_1 Sex_{female} + \beta_2 GA_{<28} + \beta_3 GA_{28-32} + \beta_4 Motherage_{24-35} + \beta_5 Motherage_{35+} + \beta_4 Birthweight_{1500-2500} + \beta_4 Birthweight_{2500+} + \beta_5 Modedelevery_{forceps} + \beta_5 Modedelevery_{vacuum} + \beta_5 Modedelevery_{Ass.breech} + \beta_5 Modedelevery_{C-section} + \beta_6 priorHD_{yes} + \beta_7 Antibiotics_{yes} + \beta_8 pregnancy_{twin} + \beta_8 pregnancy_{triplet} + \beta_9 antenatalcare_{yes} + \beta_{10} CardiacD_{yes} + \beta_{11} DiabetsM_{yes} + \boldsymbol{b_j} + \boldsymbol{b_{ij}} \quad (\mathbf{4.3})$$

where b_j are variations across hospitals and b_{ij} child variation within hospitals.

In order to decide which of the two random effects models, two models were fitted: one the saturated model above with two random intercepts to estimate between and within hospital variations and the other with one random intercept model to estimate between hospitals variations. AIC and likelihood ratio test (LRT) were used to compare the two models to select the better of the two models.

Table 4.5: Information criteria for comparison of one and two random intercept models

Models	AIC	BIC	LogLik	Deviance	σ_W	σ_B	P
Model with one random intercept	14372	14468	-7172.9	14346	0.7452		
Model with two	14367.4	14471.2	-7169.7	14339.4	0.7276	0.1054	0.000

random intercepts

The model with one random intercept describes within-hospital variation and the model with two random intercept indicated that child within-hospital and between-hospital variation, and σ_W and σ_B are within-hospital and between-hospital standard deviation respectively. Table 4.5, indicate that the AIC of model with two random intercept reduced from 14,372 to 14,367.4 and the deviance was reduced from 14,346 to 14,339.4.The small p-value of the log likelihood ratio test ($P < 0.001$) also indicates that the model with two random intercept is parsimonious model. P is the p-value of the log likelihood ratio test of the two models. Also when the model without random effects (i.e. simply the generalized linear model) was considered, it gives AIC value of 14,741 which is large as compared to the above two models with random effects.

Next, the covariates for the fixed effect were assessed and the candidate covariates were selected by removing covariates starting from with highest p-value (greater than 0.05) sequentially. The first covariate removed was antenatal care follow up with the highest p-value 0.929 and refitted the reduced model with the remaining covariates. The AIC is reduced from 14367.4 to 14365 and the p-value of log likelihood ratio test (p=0.938) supporting that the reduced model is the preferable one. The next variable removed is mode of delivery with p-value (p=0.920) and refitted the reduced model. The AIC is reduced from 14365 to 14363 and the p-value of log likelihood ratio test (p=0.636) supporting that the reduced model is preferable. The next variable removed was antibiotics prior to delivery with the highest p-value 0.639 and refitted the reduced model with the remaining covariates. The AIC is reduced from 14363 to 14362 and the p-value of log likelihood ratio test (p=0.410) supporting that the reduced model is preferable one. The next variable removed maternal age at delivery with the highest p-value 0.416 and refitted the reduced model with the remaining covariates. The AIC is reduced from 14362 to 14360 and the p-value of log likelihood ratio test (p=0.244) supporting that the reduced model is preferable one.

The next variable removed was prior hypertensive disorder with the highest p-value 0.321 and refitted the reduced model with the remaining covariates. The AIC is similar with the previously reduced model but still the log likelihood ratio test indicates that the reduced model is better with p-value (p=0.218). In addition, the

model with small number of covariates is the reduced model is to be preferred. The next variable removed is cardiac disease with p-value (p=0.103) and refitted the reduced model. For this model the AIC is reduced from 14360 to 14359 and the p-value of log likelihood ratio test (p=0.078) suggesting that the reduced model is preferable one. Therefore, the final proposed GLMM for respiratory rate is given as:

$$logit(\pi_{ij}) = \beta_0 + \beta_1 Sex_{female} + \beta_2 GA_{28-32} + \beta_3 GA_{less\ than\ 28} + \beta_4 Birthweight_{1500-2500} + \beta_5 Birthweight_{>2500} + \beta_6 DiabetecD_{yes} + \boldsymbol{b_j} + \boldsymbol{b_{ij}} \quad (\mathbf{4.4})$$

Table 4.6: Parameter estimates (standard errors) and corresponding P value for GLMM based on Model (4.4)

Factors	Level	Estimates ($\hat{\beta}$)	Standard Error	p-value	aOR	95% CI aOR
Intercept		-0.084	0.332	0.801	0.920	0.479, 1.761
Infant Sex	Male(ref)	..			1	
	Female	-0.148	0.063	0.019	0.862	0.760, 0.974
Gestational age in weeks.	Moderate preterm (ref)	..			1	
	Very preterm	0.511	0.149	0.001	1.667	1.243, 2.234
	Extremely preterm	0.677	0.155	0.000	1.968	1.453, 2.664
Birth weight in grams	Less than 1500 (ref)	..			1	
	Between 1500-2500	-0.330	0.072	0.000	0.719	0.625, 0827
	Greater than 2500	-0.353	0.183	0.053	0.702	0.491, 1.001
Diabetic	No (ref)	..			1	

	Yes	0.696	0.299	0.019	1.990	1.117, 3.60
$\sigma^2{}_W$ = 0.529 $\sigma^2{}_B$ = 0.011						

Ref=reference category

Unlike the marginal model, in the generalized estimating equation where parameters are treated as population averages, in the GLMM analysis, parameter interpretation is based on specific subjects or hospitals. The parameter interpretation is conditional on the random effects, which is common to all individual children in the same hospital.

Given the same random intercept $b_{j,}$ the estimated odds-ratio of abnormal respiratory rate of female preterm infants were (aOR=0.862, 95% CI: 0.760, 0.974) 14% less likely than male preterm infants in the same j^{th} hospital keeping constant the other fixed effect variable in the model.

At the given constant random effect, preterm infants who very preterm were 66% more likely to have abnormal respiratory rate than moderate preterm (aOR=1.667, 95% CI: 1.243, 2.234). Similarly, a preterm infant who extremely very preterm were 97% more likely abnormal respiratory rate than who are moderate preterm (aOR=1.968, 95% CI: 1.453, 2.664).

At the given constant random effect, in the same way the estimated odds that a child born who are birth weight between 1500-2500 were 18% less likely to have abnormal respiratory rate than preterm infants who's their birth weight less than 1500grams in the same hospital (aOR=0.72, 95% CI: 0.625, 0827).

At the given constant random effect, the odds that a preterm infants born to mother who from diabetic disease is (aOR=1.99, 95% CI: 1.117, 3.60) times higher to have abnormal respiratory rate compared to one whose mother is not diabetic disease. This shows that mothers deliver child with abnormal respiratory rate for mothers who are diabetic were 99% more likely abnormal respiratory rate than mothers who are not diabetic disease.

4.5 Model diagnostic for GLMM

Residuals versus observation cluster ID number plot panel one, suggested that the residuals are approximately symmetric around zero. Q-Q plots for normality of random effects at hospital and individual levels are also given in the figure at panel two and three, and illustrates that the random effects are normally distributed with mean zero and variance covariance matrix D. Thus, the fitted GLMM model is fine for the given data.

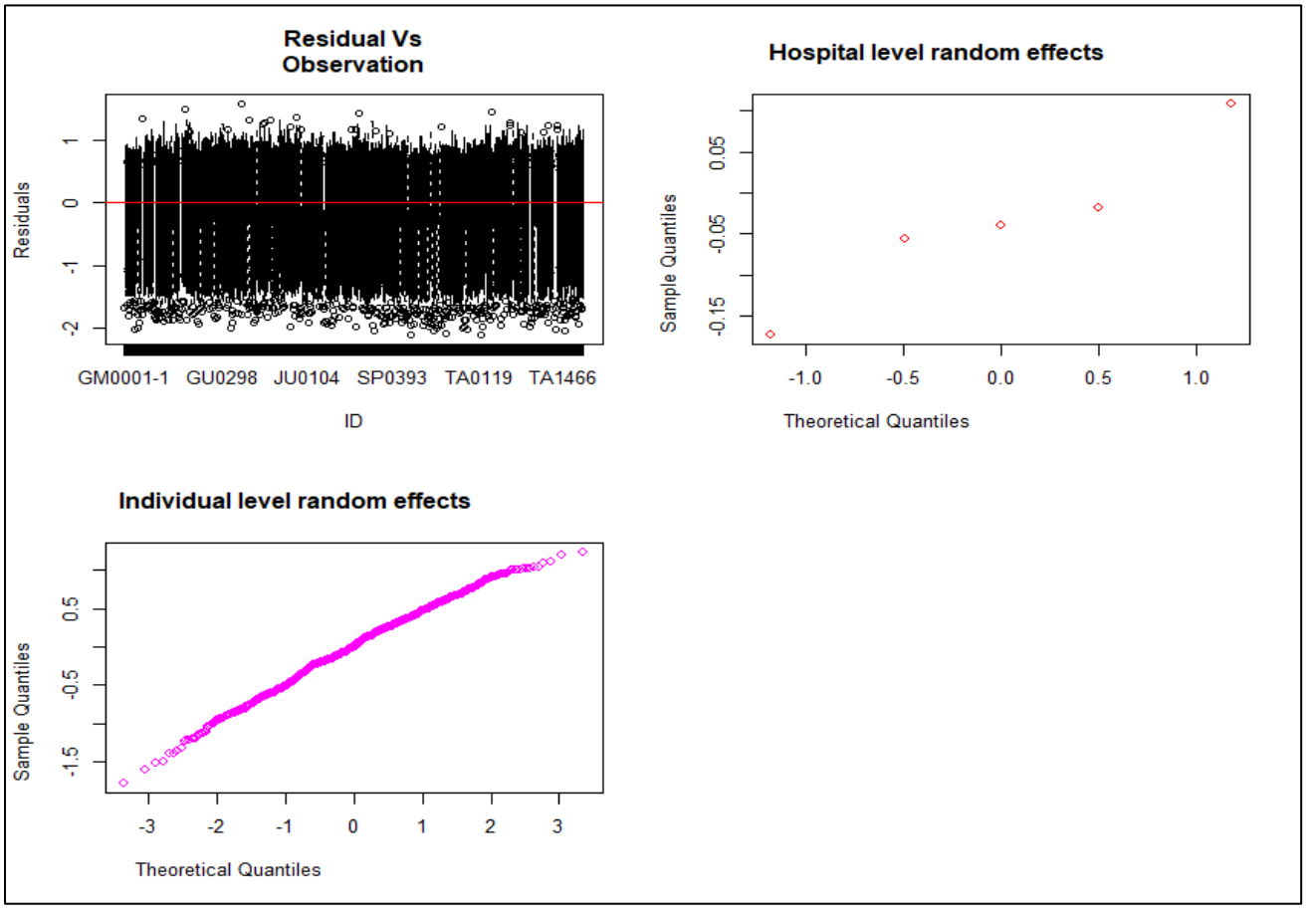

Figure 4.4: Diagnosis plots for the generalized linear mixed model

CHAPTER FIVE

5. Discussion and conclusion

5.1 Discussion

This study was aimed at modeling the determinants of respiratory rate problem in Ethiopia. In this section the result of the current study will be discussed by relating results of the previous study, i.e. Antonio & Beirlant, (2007), Altman *et al.*, (2018), Wen *et a*l., (2019), Zanardo et al., (2009) and Anna *et a*l., (2018). It should be noted that there is inconsistency in the conclusion from the analysis of various methodology, which might be due to the fact that they make use of varying amount of information, which determines the power of their inferences. Thus, the analysis was extended to other statistical methods to account for the clustered nature of correlated observations. The data were then analyzed using two model families one with marginal models (GEE) and the other is generalized linear mixed model (GLMM).

Four proposed working correlation structures; exchangeable, independence, unstructured and autoregressive correlation assumptions were taken for comparison in generalized estimating equation (GEE) model-building strategy. The model with exchangeable working correlation structure was found to give a better fit to the data than independence; unstructured and autoregressive. This supports that considered the clustering nature of the data was essential for the analysis and the dependency of individuals for the given data.

The purpose of GLMM was to evaluate within-hospital and between-hospital variations of respiratory rate problem in Ethiopia. Two models were fitted: one with only one random intercept model to assess only within-hospital variation and another with two random intercepts model, in order to account within-hospital and between-hospital variations.

Additionally, generalized linear model (GLM) was fitted for the sake of comparison whether including random effects in the analysis is important or not. The three models were compared using the AIC value followed by likelihood ratio test and we got a model with two random intercepts was favorable. This demonstrates that, accounting for within-hospital and between-hospitals variations for the analysis of respiratory rate to show within-hospital and between-hospital heterogeneity in respiratory rate. This finding is supported by the explanation or su*g*gestion of Antonio & Beirlant, (2007).

Both fitted models lead to the same conclusion that infant sex, gestational age and birth weights were significantly associated with respiratory rate. Also maternal with diabetic mellitus disease significantly associated with respiratory rate in GLMM fitted model. This study found that female gender has a protective effect against respiratory rate. A female child is less likely to be born with abnormal respiratory rate than male child. This agrees with study of Altman *et al.*, (2018), study of Wen *et a*l., (2019) and study of Anna *et a*l., (2018)

This study shows a negative association between gestational age in weeks and abnormal respiratory rate which agreeing with the findings in Italy Zanardo, et al., (2009) and Sweden Altman *et al., (*2018). The study shows that the odds of gestational age with abnormal respiratory rate consistently decreased as the gestational age in weeks increased. One of the most predominant causes of abnormal respiratory rate is the gestational age. The chance of having abnormal respiratory rate is higher among extremely preterm child born. This is similar with finding of Zanardo, et al., (2009), Wen *et a*l., (2019) and Altman *et al., (*2018).

There was also a significant association between abnormal respiratory rate and child birth-weight. According to this study, birth-weight emerged as a strong determinant for respiratory rate. The odds of having infants with abnormal respiratory rate were higher among preterm infants children birth weight less than 1500g. A low child birth weight increased the risk of having a respiratory rate. These findings of this study are similar to a study done in Italy by Zanardo et al., (2009).

5.2 Conclusion

In conclusion, the present study showed that infant sex, gestational age, and birth weight were statistically significant factors of abnormal respiratory rate. Therefore, giving awareness on feeding style and motivating pregnant women for having antenatal care is highly recommended in order to improve the birth weight of infants.

References

Agresti, A., 2002. Categorical data analysis John Wiley & sons. *Inc., Publication*, *15*, p.24.

Alemayehu, S., Kidanu, K., Kahsay, T. and Kassa, M., 2019. Risk factors of acute respiratory infections among under five children attending public hospitals in southern Tigray, Ethiopia, 2016/2017. *BMC pediatrics*, *19*(1), p.380.

Altman, M., Vanpée, M., Cnattingius, S. and Norman, M., 2013. Risk factors for acute respiratory morbidity in moderately preterm infants. *Paediatric and perinatal epidemiology*, *27*(2), pp.172-181.

Anna Niesłuchowska-Hoxha., Cnota, W., Czuba, B., Ruci, A., Ciaciura-Jarno, M., Jagielska, A., Wójtowicz, D., Kierach, R., Dąbrowski, K., Sidorowicz, M. and Skrzypulec-Plinta, W., 2018. A Retrospective Study on the Risk of Respiratory Distress Syndrome in Singleton Pregnancies with Preterm Premature Rupture of Membranes between 24. *BioMed research international*, *2018*.

Antonio, K. and Beirlant, J., 2007. Actuarial statistics with generalized linear mixed models. *Insurance: Mathematics and Economics*, *40*(1), pp.58-76.

Been, J.V., Lugtenberg, M.J., Smets, E., van Schayck, C.P., Kramer, B.W., Mommers, M. and Sheikh, A., 2014. Preterm birth and childhood wheezing disorders: a systematic review and meta-analysis. *PLoS Med*, *11*(1), p.e1001596.

Belsley, D.A., Kuh. E., 8: Welsch, RE 1980. Regression diagnostics: Identifying influential data and sources of collinearity.

Blencowe, H., Cousens, S., Oestergaard, M.Z., Chou, D., Moller, A.B., Narwal, R., Adler, A., Garcia, C.V., Rohde, S., Say, L. and Lawn, J.E., 2012. National, regional, and worldwide estimates of preterm birth rates in the year 2010 with time trends since 1990 for selected countries: a systematic analysis and implications. *The lancet*, *379*(9832), pp.2162-2172.

Booth, J.G. and Hobert, J.P., 1999. Maximizing generalized linear mixed model likelihoods with an automated Monte Carlo EM algorithm. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *61*(1), pp.265-285

Cook, R.D. and Weisberg, S., 1982. *Residuals and influence in regression*. New York: Chapman and Hall.

Davison, A.C. and Snell, E.J., 1991. Residuals and diagnostics. *Statistical Theory and Mod*.

Debelew, G.T., Afework, M.F. and Yalew, A.W., 2014. Determinants and causes of neonatal mortality in Jimma zone, southwest Ethiopia: a multilevel analysis of prospective follow up study. *Plos one*, *9*(9), p.e107184.

Diggle, P.J., Liang¡. Y. and Zeger SL (1994). Analysis of Longitudinal Data. *Biometrika, Vol. 80, pp. 517-526*

Edwards, M.O., Kotecha, S.J., Lowe, J., Richards, L., Watkins, W.J. and Kotecha, S., 2016. Management of prematurity-associated wheeze and its association with atopy. *PloS one*, *11*(5), p.e0155695.

Hibbard, J.U., Wilkins, I., Sun, L., Gregory, K., Haberman, S., Hoffman, M., Kominiarek, M.A., Reddy, U., Bailit, J., Branch, D.W. and Burkman, R., 2010. Respiratory morbidity in late preterm births. *JAMA: the journal of the American Medical Association*, *304*(4), p.419.

Hooker, E.A., O'Brien, D.J., Danzl, D.F., Barefoot, J.A. and Brown, J.E., 1989. Respiratory rates in emergency department patients. *The Journal of emergency medicine*, *7*(2), pp.129-132.

Liang, K.Y. and Zeger, S.L., 1986. Longitudinal data analysis using generalized linear models. *Biometrika*, *73*(1), pp.13-22.

McCullagh, P., 81. Nelder. JA (1989), Generalized Linear Models. *CRC Monographs on Statistics & Applied Probability, Springer Verlag, New York*.McCullagh, P., 1983. Quasi-likelihood functions. *The Annals of Statistics*, pp.59-67.

Molenberghs, G., & Verbeke, G., (2005).Models for Discrete Longitudinal Data. Library of Congress

Muhe, L.M., McClure, E.M., Nigussie, A.K., Mekasha, A., Worku, B., Worku, A., Demtse, A., Eshetu, B., Tigabu, Z., Gizaw, M.A. and Workneh, N., 2019. Major causes of death in preterm infants in selected hospitals in Ethiopia (SIP): a prospective, cross-sectional, observational study. *The Lancet Global Health*, *7*(8), pp.e1130-e1138.

Myers, R.H., Montgomery, D.C., Vining, G.G. and Robinson, T.J., 2010. Generalized Linear Models with Applications in Engineering and the Sciences. A John Wiley and Sons. *Inc. New Jersey*, pp.202-264.

National Collaborating Centre for Women's and Children's Health (UK. "Feverish illness in children: assessment and initial management in children younger than 5 years." (2013).

Northway Jr, W.H., Rosan, R.C. and Porter, D.Y., 1967. Pulmonary disease following respirator therapy of hyaline-membrane disease: bronchopulmonary dysplasia. *New England Journal of Medicine*, *276*(7), pp.357-368.

Pan, W., 2001. Akaike's information criterion in generalized estimating equations. *Biometrics*, *57*(1), pp.120-125.

Petersen, T., 2004. *Analyzing panel data: Fixed-and random-effects models* (pp. 331-345). na.

Preisser, J.S. and Qaqish, B.F., 1996. Deletion diagnostics for generalised estimating equations. *Biometrika*, *83*(3), pp.551-562.

Wen, Y.H., Yang, H.I., Chou, H.C., Chen, C.Y., Hsieh, W.S., Tsou, K.I. and Tsao, P.N., 2019. Association of Maternal Preeclampsia with Neonatal Respiratory Distress Syndrome in Very-Low-Birth-Weight Infants. *Scientific reports*, *9*(1), pp.1-8.

World Health Organization, 1997. The world health report 1997: conquering suffering; enriching humanity.

World Health Organization, 2014. Every newborn: an action plan to end preventable deaths. *Geneva: World Health Organization*.

Zanardo, V., Vedovato, S., Cosmi, E., Litta, P., Cavallin, F., Trevisanuto, D. and Chiarelli, S., 2010. Preterm premature rupture of membranes, chorioamnion inflammatory scores and neonatal respiratory outcome. *BJOG: An International Journal of Obstetrics & Gynaecology*, *117*(1), pp.94-98.

Zeger, S.L. and Karim, M.R., 1991. Generalized linear models with random effects; a Gibbs sampling approach. *Journal of the American statistical association*, *86*(413), pp.79-86.

Zorn, C.J., 2001. Generalized estimating equation models for correlated data: A review with applications. *American Journal of Political Science*, pp.470-490.

Appendix

Table 4.3 The four proposed working correlation structure for empirical and model-based standard error

	Exchangeable			Independent		
		Model-based	Empirical		Model-based	Empirical
Coeff.	**Estimates**	**(S.E)**	**(S.E)**	**Estimates**	**(S.E)**	**(S.E)**
β_0	-0.434	0.375	0.560	-0.049	0.275	0.563
β_1	-0.143	0.056	0.070	-0.119	0.041	0.063
β_2	0.444	0.130	0.137	0.383	0.093	0.136
β_3	0.594	0.134	0.146	0.458	0.096	0.140
β_4	0.035	0.060	0.077	0.000	0.044	0.069
β_5	0.023	0.099	0.111	-0.029	0.069	0.110
β_6	-0.302	0.063	0.101	-0.264	0.046	0.077
β_7	-0.328	0.160	0.183	-0.266	0.121	0.181
β_8	-1.476	0.641	1.006	-1.100	0.521	0.828
β_9	0.563	0.336	0.347	0.400	0.243	0.324
β_{10}	0.180	0.477	0.490	0.343	0.347	0.341
β_{11}	0.005	0.067	0.069	0.038	0.048	0.073
β_{12}	-0.066	0.067	0.071	-0.034	0.048	0.076
β_{13}	-0.025	0.064	0.066	-0.052	0.045	0.068
β_{14}	0.023	0.061	0.066	0.048	0.044	0.068
β_{15}	0.227	0.222	0.186	0.166	0.134	0.153
β_{16}	-0.073	0.114	0.127	-0.034	0.082	0.129
β_{17}	0.508	0.274	0.377	0.480	0.200	0.356

β_{18}	0.530	0.258	0.331	0.304	0.201	0.341

	Autoregressive Order 1 (AR1)			Unstructured		
		Model-based	Empirical		Model-based	Empirical
Coeff.	**Estimates**	**(S.E)**	**(S.E)**	**Estimates**	**(S.E)**	**(S.E)**
β_0	-0.473	0.547	0.472	-0.078	0.283	0.543
β_1	-0.188	0.079	0.066	-0.096	0.041	0.058
β_2	0.491	0.180	0.123	0.372	0.092	0.124
β_3	0.684	0.185	0.143	0.430	0.096	0.127
β_4	0.052	0.085	0.070	-0.001	0.044	0.064
β_5	0.031	0.139	0.096	-0.046	0.069	0.104
β_6	-0.554	0.084	0.126	-0.327	0.047	0.071
β_7	-0.499	0.227	0.175	-0.308	0.126	0.172
β_8	-1.949	1.028	1.384	-1.094	0.552	0.818
β_9	0.526	0.471	0.358	0.408	0.256	0.305
β_{10}	-0.006	0.654	0.406	0.297	0.361	0.327
β_{11}	-0.017	0.095	0.057	0.030	0.048	0.069
β_{12}	-0.038	0.095	0.060	-0.034	0.049	0.073
β_{13}	-0.008	0.090	0.053	-0.038	0.046	0.063
β_{14}	0.035	0.086	0.056	0.072	0.044	0.066
β_{15}	0.083	0.280	0.117	0.329	0.057	0.076

β_{16}	-0.039	0.163	0.108	-0.009	0.083	0.125
β_{17}	0.419	0.398	0.314	0.515	0.203	0.337
β_{18}	0.492	0.382	0.283	0.279	0.208	0.326

Table 4.4 Variable selections in GEE for the full model empirical Standard Error Estimates

Parameter	**Estimate**	**Standard Error**	**95% Confidence Limits**		**Z**	**Pr > |Z|**
β0	-0.434	0.560	-1.531	0.663	-0.78	0.438
β1	-0.143	0.070	-0.280	-0.006	-2.04	0.042
β2	0.444	0.137	0.177	0.712	3.25	0.001
β3	0.594	0.146	0.307	0.880	4.06	<.000
β4	0.035	0.077	-0.116	0.185	0.45	0.649
β5	0.023	0.111	-0.195	0.240	0.2	0.839
β6	-0.302	0.101	-0.499	-0.105	-3	0.003
β7	-0.328	0.183	-0.687	0.030	-1.8	0.073
β8	-1.476	1.006	-3.447	0.495	-1.47	0.142
β9	0.563	0.347	-0.116	1.242	1.63	0.104

β10	0.180	0.490	-0.780	1.140	0.37	0.713
β11	0.005	0.069	-0.130	0.140	0.07	0.946
β12	-0.066	0.071	-0.205	0.074	-0.92	0.359
β13	-0.025	0.066	-0.154	0.105	-0.37	0.708
β14	0.023	0.066	-0.107	0.152	0.34	0.731
β15	0.227	0.186	-0.136	0.591	1.23	0.221
β16	-0.073	0.127	-0.321	0.175	-0.58	0.564
β17	0.508	0.377	-0.231	1.248	1.35	0.178
β18	0.530	0.331	-0.119	1.178	1.6	0.110