

**TULANE UNIVERSITY LIBRARY
HOWARD-TILTON MEMORIAL LIBRARY**

Manuscript Theses

Unpublished theses submitted for the Honors, Master's and Doctor's degrees and deposited in the Howard-Tilton Memorial Library may be inspected, but are to be used only with due regard to the rights of the authors. Bibliographical references may be noted, but passages may be copied or closely paraphrased only with the written permission of the author, and proper credit must be given in subsequent written or published work.

This thesis by Chansatitporn, Natkamol has been used by the following persons, whose signatures attest their acceptance of the foregoing restrictions.

SIGNATURE

ADDRESS

DATE

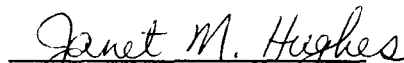
**COMPARISON OF GENERALIZED ESTIMATING EQUATIONS AND
RANDOM EFFECTS MODELS FOR LONGITUDINAL BINARY OUTCOMES:
APPLICATION TO SPEECH DELAY IN THAI CHILDREN**

**A DISSERTATION
SUBMITTED ON THE SEVENTEENTH DAY OF DECEMBER 2009
TO THE DEPARTMENT OF BIostatISTICS
IN PARTIAL FULFILLMENT OF THE REQUIRMENTS
OF THE SCHOOL OF PUBLIC HEALTH AND TROPICAL MEDICINE
OF TULANE UNIVERSITY
FOR THE DEGREE
OF
DOCTOR OF SCIENCE
BY**


Natkamol Chansatitporn
Natkamol Chansatitporn

APPROVED: 
Janet C. Rice, Ph.D.
Committee Chair


John J. Lefante, Jr., Ph.D.


Janet M. Hughes, Ph.D.


Nichara Ruangdaraganon MD.

UMI Number: DP18863

All rights reserved

INFORMATION TO ALL USERS

The quality of this reproduction is dependent on the quality of the copy submitted.

In the unlikely event that the author did not send a complete manuscript and there are missing pages, these will be noted. Also, if material had to be removed, a note will indicate the deletion.



UMI DP18863

Copyright 2012 by ProQuest LLC.

All rights reserved. This edition of the work is protected against unauthorized copying under Title 17, United States Code.



ProQuest LLC.
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106 - 1346

ABSTRACT

Background: The analysis of data from longitudinal studies requires special techniques which can handle the correlation between repeated measurements in the same subjects. Marginal (Generalized Estimating Equations: [GEE]) and subject-specific models (Random Effects models: [RE]) are two methods used to analyze correlated data.

Purpose: To compare logistic regression results using GEE and RE models regarding regression coefficients and standard errors. The results of these two techniques were also compared to those from ordinary logistic regression [OLR] in which the time dependency is ignored. We applied these techniques to the identification of risk factors associated with suspected speech delay (SSD) in Thai children.

Methods: Longitudinal data of child language development in 4,245 Thai children were considered. Two data sets were derived from the full data: complete data (children with no missing value) and complete at baseline data (children with complete data and children with one missing observation on the repeated outcomes at 18 or 24 months). OLR, logistic regression using GEE and RE models were applied to both data sets to model binary outcomes of SSD at three time points (12 months [baseline], 18 and 24 months).

Results: The regression coefficients from RE models are larger than those from both OLR and GEE models. The standard errors obtained from RE models are also larger than those from the OLR and GEE modes. The statistical tests from all three approaches are similar, however in the OLR models: the standard errors are over-estimated for time-varying covariates and under-estimated for time-invariant covariates. The results of

analyses from both data sets indicated the same set of risk factors for SSD, namely age, gender, low birth weight and father's education; however, the direction of the association for father's education is not consistent.

Conclusions: The choice between GEE and RE models for analyzing correlated data depends mainly on the aim of the study. The GEE approach is appropriate for making inferences on the population, and the RE model is suitable for making inferences for an individual. Restricting the analysis to the complete cases may lead to biased parameter estimates, especially when a large percentage of data are missing.

ACKNOWLEDGEMENTS

I would first like to express my endless gratitude to Dr. Janet Rice, chairman of dissertation committee, for the invaluable guidance and advice. Your help and encouragement enabled me to finally accomplish my dissertation.

I would like to express my forever gratitude to Dr. Janet Hughes, a member of my dissertation committee; she gave not only the guidance and advice on my dissertation, but also continued great support to me from the beginning to the end of my study.

I would like to extend my appreciation to Dr. John Lefante for serving on my dissertation committee and providing valuable comments on my dissertation.

I am deeply grateful to Dr. Nichara Ruangdaraganon, a member of my dissertation committee, for her helpful comments on child language development and for her trust in allowing me to utilize the PCTC study data for my dissertation.

Finally, I would like to express my deepest gratitude to my family: my mother, sisters and brothers for their unconditional love, support and encouragement. Without all of them, I would not have been able to complete this study.

TABLE OF CONTENTS

ABSTRACT	i
ACKNOWLEDGEMENT	ii
TABLE OF CONTENTS	iii
LIST OF TABLES	iv
CHAPTER I. INTRODUCTION	1
1.1 Background and Significant of the Study	1
1.2 Research Questions	3
1.3 Objectives	4
CHAPTER II. REVIEW OF LITERATURE	5
2.1 Ordinary Logistic Regression (OLR)	5
2.2 Generalized Estimating Equations (GEE)	7
2.4 Random Effects Models	12
2.5 Interpretation and Relationship between the Regression Coefficients from Population-averaged Approach and Subject-specific Approach	16
2.6 Literature Review of the Statistical Properties of OLR,GEE and RE models	18
2.7 Review of Speech and Language Development	31
2.8 Risk Factors Related to Speech and Language Development	35
CHAPTER III. METHODOLOGY	47
3.1 Dataset	47
3.2 Outcome and Covariates	48
3.3 Statistical Models	49

3.4 Statistical Analysis	50
3.5 Missing data	52
CHAPTER IV RESULTS	53
4.1 Description of Children in the PCTC Data	54
4.2 Longitudinal Data Modeling for Predicting SSD Based on Complete Data	60
4.2.1 Fitting Ordinary Logistic Regression Model Based on Complete Data	63
4.2.2 Fitting GEE Logistic Regression Models Based on Complete Data	65
4.2.3 Fitting Random Effects Logistic Models Based on Complete Data	68
4.2.4 Comparison of Models Derived from OLR, GEE and RE Models Based on Complete Data	71
4.3 Longitudinal Data Modeling for Predicting SSD Based on Complete at Baseline Data	75
4.3.1 Fitting Ordinary Logistic Regression Model Based on Complete at Baseline Data	77
4.3.2 Fitting GEE Logistic Regression Models Based on Complete at Baseline Data	78
4.3.3 Fitting Random Effect Logistic Regression Models Based on Complete at Baseline Data	80
4.3.4 Comparison of Models Derived from OLR, GEE and RE models Based on Complete at Baseline Data	82
4.4 Comparison Results from Complete Data and Complete at Baseline Data	85
CHAPTER V DISCUSSION AND CONCLUSIONS	86
5.1 Identification of Factors Associated with Suspected Speech Delay	86
5.2 Comparison between Marginal and Subject-specific Models	88
5.3 Relationship between Marginal and Subject-specific Models	91
5.4 Interpretation of the Regression Coefficients	92
5.5 Impact of Missing Data on Regression Inference	92

5.6 Conclusions	94
APPENDIX A	96
APPENDIX B	97
REFERENCES	101

LIST OF TABLES

Table 1	Summary of the statistical properties of ordinary logistic regression, generalized estimating equations and random effects models based on theories	25
Table 2	Summary of the statistical properties of ordinary logistic regression, generalized estimating equations and random effects models based on simulation studies.	26
Table 3	Summary of the statistical properties of ordinary logistic regression, generalized estimating equations and random effects models based on empirical studies ..	28
Table 4	Normal speech and language development milestones	33
Table 5	Summary of studies evaluating risk factors affecting speech and language Development	42
Table 6	Number and percent of missing data for each variables used in analysis	55
Table 7	Comparison between complete and incomplete data on each covariates	56
Table 8	The distribution of baseline characteristics of the birth cohort based on complete data (N=1,823).....	57
Table 9	Summary statistics on the result of speech delay assessment across three time points of follow-up (Complete data, N =1,823)	59
Table 10	Number and percentage of SSD cases at each time point (Complete data).....	59
Table 11	Results of bivariate analysis using ordinary logistic regression based on complete data	62
Table 12	Main effects logistic regression model predicting SSD for complete data	64
Table 13	Results of testing interaction terms to be added in main effect logistic regression model for complete data	65
Table 14	GEE logistic model (exchangeable) predicting SSD for complete data	66
Table 15	GEE logistic model (autoregressive) predicting SSD for complete data	66
Table 16	Random effect logistic model (random intercept) predicting SSD for complete data	69

Table 17	Random effect logistic model (random intercept and slope) predicting SSD for complete data	69
Table 18	Regression coefficients and standard errors obtained from fitting marginal and random effects logistic regression models to predict SSD in Thai children based on complete data	74
Table 19	Results of bivariate analysis using ordinary logistic regression for complete at baseline data	76
Table 20	Logistic regression model predicting SSD for complete at baseline data	77
Table 21	GEE logistic model (exchangeable) predicting SSD for complete at baseline data	79
Table 22	GEE logistic model (autoregressive) predicting SSD for complete at baseline data	79
Table 23	Random effect logistic model (random intercept) predicting SSD for incomplete data	81
Table 24	Random effect logistic model (random intercept and slope) predicting SSD for incomplete data	81
Table 25	Regression coefficients and standard errors obtained from fitting marginal and random effects logistic regression models to predict SSD in Thai children based on complete at baseline data	84

LIST OF FIGURES

Figure 1 Proportions of children defined as having SSD at each age group 61

CHAPTER I

INTRODUCTION

1.1 Background and Significance of the Study

Longitudinal studies are common in many fields of health sciences research (Campbell, 2000; John M. Neuhaus, 2001). They are defined as studies in which subjects are measured repeatedly over multiple times. This type of study allows investigators to assess how each subject's response changes under different experimental settings. The ability to relate changes in explanatory variables for a given subject to change in the outcomes of that same subject distinguishes longitudinal studies from cross-sectional studies. With cross-sectional data, one can only measure difference of outcome between groups or units classified by the explanatory variables, but can not draw conclusions about how a particular subject would change over time.

One of the essential assumptions of the ordinary least square (OLS) method in regression analysis or classic logistic regression is that residuals are independent. This situation would occur when every observation in the study is independent of the others. This assumption is not satisfied under longitudinal studies in which repeated measurements are taken over time on the same subjects. It is well recognized that repeated observations of the same subject tend to be more alike than measurements obtained from different subjects. Using ordinary logistic regression to analyze data from longitudinal data results in incorrect variance estimates and inefficient estimates of the regression parameters (Heo & Leon, 2005; Peter, Richard, Bankhead, Ades, & Strene, 2003; Ukoumunne, Carlin, & Gulliford, 2007; Zeger & Liang, 1992). Analysis of such

correlated data, therefore, needs to accommodate the statistical dependence or correlation among the repeated measurements within subjects in order to obtain valid inference.

In longitudinal analysis with binary outcomes, building statistical models to evaluate changes over time and the effect of explanatory variables is more complicated than for continuous outcomes (Carlin, Wolf, Brown, & Gelman, 2001; J.M. Neuhaus, Kalbfleisch, & Hauch, 1992). Several approaches have been developed and proposed to take account of within-subject correlation. Most of these approaches can be broadly classified into two groups: marginal or population-averaged and subject-specific models (Zeger, Liang, & Albert, 1988). In the marginal model, the expectation of the outcome is modeled as a function of the covariates and the dependence of observations is taken into account by assuming a working correlation structure for the repeated measurements among the outcome (Liang & Zeger, 1986; Zeger & Liang, 1986). The estimated regression parameters have interpretation for the population rather than for any individual. Examples of this approach include the generalized estimating equations (GEE) method (Liang & Zeger, 1986) and the beta-binomial regression model (Prentice, 1986). In the subject-specific model, the heterogeneity across subjects is explicitly modeled; the probability of a binary outcome is modeled as a function of the covariates and specific parameters of the individual subjects. The estimated regression coefficients represent effects specific to the subject or condition on the value of the random effects. Examples of this approach include the random effects models (RE) (Stiratelli, Laird, & Ware, 1984) or multilevel and the conditional likelihood approach for matched pair data.

In recent years the computational complexity of fitting models to binary data has been overcome. In particular, procedures in statistical software such as SAS and STATA

have become available to fit the logistic random effects model and the marginal model using GEE's. As the GEE and RE models are alternative methods for the analysis of binary repeated observations, an understanding of the relationships between parameter estimates from these methods is essential in choosing an appropriate analysis method. In general, random effects models and GEE's handle the statistical dependence of the repeated data differently, and consequently can lead to different parameter estimates (Zeger, Liang, & Albert, 1988; Zeger, Liang, & Albert, 1991). Although comparisons of the two methods have appeared in the statistical literature, there have been a limited number of studies for binary outcomes.

The purpose of this study is to compare the regression parameter estimates and standard errors obtained from GEE logistic model and random-effects logistic model by analyzing a longitudinal child language development dataset from Thailand with three time points. The results of these two methods will also be compared to those obtained from ordinary logistic regression in which the time dependency is ignored.

1.2 Research Questions

1. How do regression parameter estimates and their standard errors obtained from logistic regression using GEE and random effects models in the context of longitudinal binary data analysis differ?
2. What risk factors are associated with suspected speech delay among Thai children at three time points, 12, 18 and 24 months of age?

1.3 Objectives

1. To compare regression parameter estimates and their standard errors from logistic regression using GEE and random effects models in the context of longitudinal binary data analysis.
2. To determine risk factors associated with suspected speech delay in Thai children.

CHAPTER II

LITERATURE REVIEW

The literature review has two sections. The first section discusses the statistical properties of the analytic methods. This section begins with describing standard logistic regression analysis which assumes independence. It describes the impact of violating the assumption of independence on the parameter estimates and their standard errors. Next is a discussion of two alternative statistical approaches: the GEE method, representative of the population-average approach, and random effects models, representative of the subject-specific approach. This section reviews findings from statistical theory, simulation studies and empirical studies. The second section discusses the development of speech in children as well as previous studies that assess the risk factors affecting speech development.

2.1 Ordinary Logistic Regression (OLR)

Logistic regression (Hosmer & Lameshow, 2000) is a popular technique applied to data with binary outcomes. It is a population-average approach. Under ordinary logistic regression, the logit transformation of the marginal mean response (the probability of outcome) is modeled and parameter estimates are obtained using maximum likelihood estimation under the assumption of independent observations. The method is inappropriate for longitudinal studies because neither the model nor the estimation procedure incorporates the correlation of repeated observations. The ordinary

logistic regression model is described in the context of data from a longitudinal study as follows.

$$\log \frac{\Pr(y_{ij} = 1)}{1 - \Pr(y_{ij} = 1)} = \beta_0 + \beta_1 x_{1ij} + \beta_2 x_{2ij} + \dots + \beta_k x_{kij} + \varepsilon_{ij} \quad (1)$$

where,

y_{ij} = the binary outcome for the subject i at time j

β_0 = the constant baseline log odds

$\beta_1, \beta_2, \dots, \beta_k$ = the log odds ratio corresponding to a 1-unit increase in covariate k

$x_{1ij}, x_{2ij}, \dots, x_{kij}$ = a set of k explanatory variables for the subject i at time j

$\varepsilon_{ij} \sim N(0, \sigma^2)$ = the random error term for subject i at time j ; assumed to be independent for all i and j

Note that j is the sequence of observations for individuals, and it ranges from 1 to J . In the same way, i is the index for individuals, and it ranges from 1 to n .

2.1.1 Impact of Ignoring Dependencies

In longitudinal studies with a binary outcome, the analysis must take account of the correlations on repeated observations to obtain valid inference about regression parameters. Ignoring correlation when it exists results in two problems: incorrect estimation of the regression coefficient variances and inefficient estimates of the regression coefficients (Diggle, Liang, & Zeger, 2002; Liang & Zeger, 1993; Zeger & Liang, 1992).

According to the impact on parameter estimates when dependency is not taken into account, the ordinary logistic regression model is inadequate to provide the valid

parameter estimates. Alternative statistical methods accounting for the dependence circumstance are required. The next two sections describe two statistical techniques that handle correlated responses: GEE and random effects models which are representatives of the population-averaged and the subject-specific approaches, respectively.

2.2 Generalized Estimating Equations (GEE)

The GEE methodology was developed by Liang and Zeger to produce more efficient regression estimates for use in analyzing longitudinal designs with nonnormal outcome variables (Liang & Zeger, 1986). The GEE method is an extension of generalized linear models (GLMs), to estimate the population averaged parameter estimates while accounting for the dependency among the repeated measurements which is a standard characteristic of longitudinal designs. This approach has become very popular, especially for analysis of categorical data (Hedeker & Gibbons, 2006). The logistic regression model using the GEE estimation technique can be written as:

$$\log \frac{\Pr(y_y = 1)}{1 - \Pr(y_y = 1)} = \beta_0 + \beta_1 x_{1y} + \beta_2 x_{2y} + \dots + \beta_k x_{ky} + \varepsilon_y \quad (2)$$

Where the terms are defined as in model (1) with the exception that the ε_y may be correlated within each subject i ; this is what distinguishes the logistic regression based on the GEE estimation technique and the ordinary logistic regression. It can be seen that the logistic regression model based on the GEE estimation technique, model (2), and the ordinary logistic regression, model (1), are the same; however, the estimation techniques

differ. The ordinary logistic regression model assumes the errors are independent and uses the maximum likelihood method to estimate parameters. In contrast, the GEE method allows the errors to be correlated and uses a quasi-likelihood method to estimate the regression coefficients and the correlation among the observations separately.

A basic feature of the GEE method is that the joint distribution of a subject's response vector y_j does not need to be specified. Instead, it is only the marginal distribution of y_j at each time point that needs to be specified. The GEE approach actually treats the time dependency as a nuisance, and a (co)variance structure or working correlation matrix for the vector of repeated observations from each subject is specified to account for the dependency among the repeated observations. The focus is on the regression of dependent variable on a set of covariates. In this regard, the GEE method yields asymptotically efficient and consistent estimation of the regression parameter vector (β) (Liang & Zeger, 1993). In other words, as the number of individuals approaches infinity, the GEE estimate of β approaches the "true" β , even if the working correlation matrix is misspecified (Zeger & Liang, 1986). This is one of the advantages of the GEE method.

2.2.1 GEE Estimation

The steps of model fitting for the GEE method can be illustrated as follows (Burton, Gurrin, & Sly, 1998):

- (i) Fit a standard logistic regression model assuming all observations to be independent.

- (ii) Take the residuals from the regression and use those to estimate the parameters which quantify the correlation between observations in the same individual.
- (iii) Refit the regression model using a modified algorithm incorporating a matrix which reflects the magnitude of correlation estimates in step (ii). At this step, new values of regression coefficients (β_0 and β_1, \dots, β_k) and ultimately new residuals are generated.
- (iv) Keep alternating between steps (ii) and (iii) until the estimates stabilize and convergence is achieved.

2.2.3 Working Correlation Structures

The working correlation matrix is required by the GEE estimation technique to estimate regression parameters. This matrix can be either specified by the researcher or estimated by the GEE method in a form that matches the expected correlation structure within the subject. The closer the assumed working correlation matrix represents the actual dependence structure of repeated measurements within a subject, the more efficient in estimating regression coefficients and the more accurate in estimating regression coefficient variances. There are several options to specify the form of the working correlation matrix. This specification will differ based on the nature of the data collected. It is recommended to specify the working correlation as accurately as possible, based on the knowledge of the longitudinal process (Albert, 1999). There are specific working correlations that are appropriate for a time-dependent correlation structure (e.g., autoregressive) and some that are not (e.g., exchangeable). For cases in which the researcher may be undecided between two structures, the measure called *quasilikelihood*

under the independence model information criterion (QIC), which is an extension of *Akaike's information criterion (AIC)* to a quasilielihood-based method (Pan, 2001), can be used to choose from competing correlation structures. The QIC score that is lowest (close to zero) is judged to be the best. The chosen working correlation, however, should make the most sense theoretically (Ballinger, 2004). Although the GEE methods are generally robust to misspecification of the correlation structure, Fitzmaurice reported that incorrect specification of the correlation structure may affect the efficiency of the parameter estimates (Fitzmaurice, 1995). This claim was supported by Ballinger, especially in cases where the specified structure does not incorporate all of the information on the correlation of measurement within subject (cluster), so that an inefficient estimator could be expected (Ballinger, 2004). The GEE approach allows the working correlation structure to be specified in a variety of ways. Four common working correlations are as follows:

1. Independence correlation structure

The simplest possible correlation structure is to assume independence. The assumption of this structure is that each observation collected from an individual is completely uncorrelated with every other observation measured; correlations are assumed to be 0 for all pair-wise combinations of the within-subject variables. If $\rho_{jj'}$ is the correlation between observation j and j' , $\rho_{jj} = 1$ and $\rho_{j'j'} = 0$.

2. Exchangeable correlation structure

This structure is also known as compound symmetry. Exchangeable assumes non-zero and uniform correlations for all pairs of within-subject variables. Under this structure, every pair of observations within an individual is assumed to be equally

correlated. Formally, $\rho_{jj} = 1$ and $\rho_{j'[j \neq j']}$ = ρ where ρ is the intraclass correlation coefficient. This is equivalent to the assumption regarding the correlation in the random effects model with fixed slopes.

3. Autoregressive correlation structure

This structure indicates that two observations taken close in time within an individual tend to be more highly correlated than two observations taken far apart in time from the same individual. As the space in time between observations increases, the correlation declines according to an exponential function of the time-lag which is determined by the user. Formally, $\rho_{jj} = 1$ and $\rho_{j'[j \neq j']}$ decrease in value as the absolute difference between j and j' gets larger. As an example, a first-order autoregressive (AR-1) correlation structure specifies that $\rho_{j'} = \rho_{|j-j'|}^2$ where ρ is the correlation when $|j-j'| = 1$.

3. Unstructured correlation structure

Unstructured assumes unconstrained pair-wise correlations where each correlation is estimated from the data. There is no assumption made about the relative magnitude of correlation between any two pairs of observations. Formally, $\rho_{jj} = 1$ and $\rho_{j'[j \neq j']}$ is free to take any value between -1 and +1. This structure is used in balanced data sets (Burton, Gurrin, & Sly, 1998) and is most efficient when there are small numbers of repeated observations for a subject (Hedeker & Gibbons, 2006).

2.4 Random Effects Models

Random effects models are also known as multi-level models. In a longitudinal study, the specific observations of the individuals are defined as the level-1 units and individuals are termed as the level-2 units. Random effects models are extensions of generalized linear models for longitudinal data. The key concept of the models is that the response is assumed to be a linear function of explanatory variables with regression coefficients that can vary from person to person. This variability represents individual's natural heterogeneity due to unmeasured variables. Instead of specifying a correlation structure explicitly as in the GEE method, random effects models instead extend a standard logistic regression model by adding random effects. In a standard logistic regression model, a regression coefficient is assumed to take the same fixed value for all individuals in a data set. In contrast, random effects are regression coefficients that are allowed to vary from individual to individual. The random effects model can be expressed as:

$$\log \frac{\Pr(y_y = 1)}{1 - \Pr(y_y = 1)} = \beta_0 + u_{0i} + (\beta_1 + u_{1i})x_{1iy} + (\beta_2 + u_{2i})x_{2iy} + \dots + (\beta_k + u_{ki})x_{kiy} + \varepsilon_y \quad (3)$$

Where,

y_y = the binary outcome for the subject i at time j

β_0 = the constant baseline log odds

u_{0i} = the individual specific deviation in the intercept (random effect),
with $u_{0i} \sim N(0, \tau_0^2)$

- u_{1i}, \dots, u_{ki} = the individual specific deviations in the slopes for person i , $u_{0i} = 0$
for covariate with fixed slope, with $u_{1i} \sim N(0, \tau_1^2), \dots, u_{ki} \sim N(0, \tau_k^2)$
- $\beta_1, \beta_2, \dots, \beta_k$ = the log odds ratio corresponding to a 1-unit increase in covariate k
- $x_{1j}, x_{2j}, \dots, x_{kj}$ = a set of k explanatory variables for the subject i at time j
- $\varepsilon_{ij} \sim N(0, \sigma^2)$ = the random error term for subject i at time j ; the ε_{ij} may be
correlated within each subject i

Under model 3, the intercept and slope are allowed to vary across individuals.

The two sources of variance for this model are:

1) Level-1 residual (within-individual) variance component (σ^2) or variance of errors: represents the variance of an individual's change over time, which is described by a regression model with a population-level intercept and slope.

2) Level-2 residual (between-individual) variance component or variance of the random effects ($\tau_0^2, \tau_1^2, \dots, \tau_k^2$): represents the variance of variation in individual intercepts and slopes.

In random effects models, each subject can have a random intercept and random slopes. The logistic model where the intercept and slope are allowed to vary across the subjects is called a “*random coefficient regression logistic model*” (model 3). If only the intercept is allowed to vary and the slopes are forced to be constant across individuals, this model is called a “*random intercept logistic model*” This is the simplest model that allows individuals to have their own intercept, but the effect of covariates on the outcome is the same for every subject. The random intercept logistic model can be written as:

$$\log \frac{\Pr(y_{ij} = 1)}{1 - \Pr(y_{ij} = 1)} = \beta_0 + u_{0i} + \beta_1 x_{1ij} + \beta_2 x_{2ij} + \dots + \beta_k x_{kij} + \varepsilon_{ij} \quad (4)$$

This model permits a separate fixed intercept for each subject. It is assumed that there is an average intercept (β_0) for the population of subjects in the study and the discrepancy (u_{0i}) between the average intercept and the true intercept in the i th subject. The u_{0i} is considered a random parameter, and assumed to have a normal distribution with mean 0 and variance τ_0^2 ; that is, $N(0, \tau_0^2)$. The random effect (u_{0i}) represents the influence of the individual on his/her repeated observations, and is not attributable to covariates. To the degree of correlation due to repeated observations on the same subject has little effect on the outcome, estimates of u_{0i} will all be near 0, and the estimate of τ_0^2 will approach 0. If, on the other hand, such correlation has a strong effect on the outcome, estimates of u_{0i} will deviate from 0 and differ for each subject i ; thus, the estimate of τ_0^2 will increase in value.

By including a random-intercept, u_{0i} , in the model, the interdependencies among the repeated observations within subjects are explicitly taken into account. It is noted that the regression coefficients in random effects models represent the effects of the explanatory variables on an individual's response. This is in contrast to the marginal model coefficients which describe the effect of explanatory variables on the population average.

2.4.1 Random Effects Models Estimation

Parameter estimation for random effects models requires sophisticated mathematical algorithms. There are three types of parameters to estimate, namely, fixed effects (β_0), level-1 random effects ($\beta_0 + u_{0i}$) and variance covariance components (τ_0^2 and σ^2) (Raudenbush & Bryk, 2002). The estimation of each parameter is dependent upon the estimation of the others. All the regression parameters are estimated iteratively since they all depend upon each other and no closed form of analytic expression is available. There are two types of maximum likelihood estimation: Maximum Likelihood (ML) and Restricted Maximum-Likelihood (REML) methods. However, REML is favored since ML can be biased downwards because it does not adjust for the degrees of freedom lost by estimating the regression coefficients. REML corrects this problem by maximizing the likelihood of sample residuals (not the sample data) and is considered superior (Diggle, Liang, & Zeger, 2002; Liao & Lipsitz, 2002). The process of model fitting consists of the following steps (Burton, Gurrin, & Sly, 1998):

- (i) Fit a standard regression model assuming all observations to be independent (equivalent to assuming $\tau_0^2 = 0$)
- (ii) Take the residuals from the regression and use these to estimate τ_0^2 and σ^2
- (iii) Refit the regression model using a modified algorithm incorporating a covariance matrix which reflects the magnitude of τ_0^2 and σ^2 and therefore takes account of the correlation structure
- (iv) Keep alternating between steps (ii) and (iii) until all estimates stabilize.

2.5 Interpretation and Relationship between the Regression Coefficients from Population-averaged Approach and Subject-specific Approach

The population-averaged or marginal approach (model based on GEE method) and the subject-specific approach (random effects models) represent two fundamentally different ways of thinking about covariate effects on the phenomenon of interest and about the nature of the correlation among observations on the same subject. It is important to note that the regression coefficients of models from these two approaches in the context of longitudinal binary outcomes have different interpretations (Kuchibhatla & Fillenbaum, 2003; Pendergast et al., 1996). For the marginal approach (GEE model), the estimated regression coefficient ($\hat{\beta}_{PA}$) represents the difference of the log odds ratio of the mean of outcome probability between two values of a covariate, where the mean is taken over all subjects and all observations, weighted by the working correlation structure used in modeling. The coefficient is the same for all individuals. Strictly speaking, $\hat{\beta}_{PA}$ is the estimated population mean log odds ratio. For subject-specific approach (random effects models), the estimated regression coefficient ($\hat{\beta}_{RE}$) corresponds to the change of the log odds ratio of the mean of outcome probability due to the covariate for a single individual with the same level on the random subject effect (u_i). Shortly stated, $\hat{\beta}_{RE}$ is the individual level log odds ratio. In general, Neuhaus et al demonstrated that when the variance of the random effects is greater than zero, then the absolute magnitude of β_{PA} is smaller than β_{RE} (J.M. Neuhaus, Kalbfleisch, & Hauch, 1991). Zeger et al showed that in a random intercept model, any regression coefficient from the population-averaged or marginal model (β_{PA}) can be approximated from the

respective regression coefficient from the subject-specific model (β_{SS}) as (Zeger, Liang, & Albert, 1988):

$$\beta_{PA} \approx \frac{\beta_{SS}}{\sqrt{1 + 0.346(\tau_0^2)}} \quad (5)$$

It can be seen that the estimate from marginal approach (GEE model) is less than or equal to the estimate from subject-specific approach (random effects models). In addition, the difference in this inequality depends on the variance of the random effect that is increasing between-subject heterogeneity which leads to increasing the difference in magnitude of these two estimates. When the variance of the random effects equals zero, the two estimates are identical.

2.6 Literature Review of the Statistical Properties of OLR, GEE and RE Models

This section reviews the findings of previous studies concerning statistical properties of the three statistical techniques. It is divided into three points: theoretical, simulation and empirical studies. The techniques are ordinary logistic regression (OLR), generalized estimating equations (GEE) and random effects models (RE). The statistical properties considered are the magnitude of the regression coefficients and their standard errors.

2.6.1 Evaluating Regression Coefficients

This section reviews the findings of previous studies exploring the characteristics of regression coefficients obtained from these three analytic methods.

Theoretical studies

In longitudinal study with binary repeated outcomes, using traditional logistic regression in which correlation is incorrectly assumed to be zero leads to the estimates of the regression coefficients which are inefficient, that is, less precise than those obtained by proper methods accounting for the correlation (Diggle, Liang, & Zeger, 2002; Liang & Zeger, 1993; Zeger & Liang, 1992). Fitzmaurice considered estimators of the logistic regression parameters in models for correlated binary responses. He demonstrated that the asymptotic relative efficiency of the estimator ($\hat{\beta}$) based on logistic regression relative to the maximum likelihood estimators of time-invariant covariates declines with increasing correlation, and the decline is most considerable when within-subject correlation is greater than 0.4. In addition, efficiency losses were large as correlation increased. The efficiency of parameters assuming independence fell to approximately

40% for the correlation of 0.5 or more. The errors are considerably large for cases in which the within-subject correlation is highly positive or highly negative. He concluded that the degree of efficiency depends on both the strength of the correlation between the responses and the type of covariates (time varying or time-invariant) (Fitzmaurice, 1995).

Zeger et al derived the relationship between regression coefficients of GEE and random effects models through the equation $\beta_{PA} \approx (0.346\tau_0^2 + 1)^{-1/2} \beta_{SS}$ when the distribution of the random effects is normal. These two approaches should have identical regression coefficients when the variance of random effect (τ_0^2) is zero (the data are not correlated) (Zeger, Liang, & Albert, 1988). In a further theoretical study, Neuhaus et al presented proofs that the equation is valid for arbitrary random effects distributions. They also showed that in cases of random intercepts and random slopes, the absolute values of the estimated coefficients for the random effects models are generally larger than those for GEE. The difference in the magnitude of the estimated coefficients increases with increasing intraclass correlation. Even though the magnitude of regression coefficients from these two methods differ, the Wald tests ($\hat{\beta} / se\hat{\beta}$) are similar (J.M. Neuhaus, Kalbfleisch, & Hauch, 1991).

Simulation studies

There is one simulation study (Heo & Leon, 2005) that evaluated the performance of four different statistical methods: OLR, GEE, two random effects models based on full likelihood, RE(FL) and penalized quasi-likelihood, RE(PQL). Regression coefficients, type I error rate, power, and standard error were evaluated across the four statistical methods through computer simulations under varying simulation parameters. The results showed that when the intraclass correlation (ICC) is not zero, the OLR and GEE_{exc}

estimates were more biased toward underestimation than the random effects estimates (the bias is computed as the average of 1000 estimates of estimated coefficient minus the pre-specified true value. Overall, regardless of the method, the bias tended to increase with the magnitude of the regression coefficients and the ICC. Their findings supported the theory that the random effects model provides larger estimated coefficients than GEE (Heo & Leon, 2005). It is noted that even though this simulation study was based on the clustered randomized controlled trial design, it is comparable to longitudinal designs in which a cluster is a subject.

Empirical studies

Hu et al compared the coefficients of GEE with exchangeable correlation structure (GEE_{exc}) and random effects models by analyzing a longitudinal smoking prevention dataset with multiple time points. They found that for all covariates, the coefficients of the random effects models were largest, the GEE_{exc} the second largest, and OLR model the smallest. The coefficients of OLR and GEE_{exc} were quite similar. Even though regression coefficients from the GEE_{exc} and random effects models were different, the Wald tests from these two approaches were similar (Hu, Goldberg, Hedeker, Flay, & Pentz, 1998). The finding of this study is consistent with the theory and was confirmed by a simulation study by Heo and Leon showing that the absolute values of the estimates from the random effects models were generally larger than those from GEE method. Many later studies listed below showed results consistent with the theory.

Zorn investigated parameter estimates of GEE and random effects models with a fixed slope for binary outcomes using data on the House of Representatives' votes on the four articles of impeachment against President Clinton. He found that the coefficients

were not substantially different across the four methods: GEE with independent correlation (GEE_{ind}), GEE with exchangeable correlation (GEE_{exc}), GEE with unstructured correlation (GEE_{unc}) and random effects models. The coefficients from the random effects models were uniformly larger than all three GEE estimates. The coefficients from the GEE_{ind} and the GEE_{exc} were identical to three decimal places. The author discussed that this is true for two reasons. First, the correlation in the exchangeable model is small (0.18), indicating only a low-to-moderate level of correlation among the votes. Second, the independent model is close to fully efficient, especially in cases where (as in the study) those covariates are not time-varying covariates. Hence, the difference between these two estimators could not be seen (Zorn, 2001).

Carriere evaluated GEE_{exc} and random effects methods (random intercept and random slope) based on cluster-randomized trials designs. The effects of both types of covariates, time-invariant and time-varying, were investigated. The results showed that the random effects model provided larger coefficients than GEE_{exc} . The differences between the estimates of the two approaches were largely dependent on the inter-individual heterogeneity which can be assessed in the random models by looking at the intercept and slope variances (Carriere & Bouyer, 2002).

Kuchibhatla et al compared GEE_{exc} and random intercept models and also included OLR to investigate the impact of ignoring correlated binary responses in longitudinal studies. The models included both time-invariants and time-varying covariates. For all covariates, the absolute estimates from the random intercept models were larger than those of both OLR and GEE_{exc} approaches, and the coefficients of OLR

could be either larger or smaller than those of GEE_{exc} (Kuchibhatla & Fillenbaum, 2003). These findings were confirmed by a later study by Ananth et al. using a large cohort of twins (a sample of 285,226 twins) data. The random effects model still gave the largest coefficients when compared to GEE_{exc} and OLR. The direction of the difference between coefficients of OLR and GEE_{exc} is not certain (Ananth, Platt, & Savitz, 2005).

2.6.2 Evaluating Standard Errors

This section reviews the findings of previous studies evaluating coefficient variance estimates or standard errors obtained from the three analytic methods.

Theoretical studies

Longitudinal analysis with binary outcomes, using OLR, which is not designed for correlated data and therefore does not account for existing time dependencies, leads to incorrect standard errors (Liang & Zeger, 1993; Zeger & Liang, 1992). The biases are dependent on whether the covariates vary with time. The OLR models tend to underestimate the standard errors of time-invariant covariates and overestimate the standard errors of time-varying covariates (Dunlop, 1994; Fitzmaurice, Laird, & Rotnitzky, 1993). When considering the magnitude of standard errors of regression coefficients obtained from the three approaches, Neuhaus demonstrated that the standard errors of random effect models are generally larger than those of both GEE and OLR. However, the Wald tests are similar (J.M. Neuhaus, Kalbfleisch, & Hauch, 1992).

Simulation studies

A simulation study by Heo et al examined the performance of four different methods: OLR, GEE and two random effects models based on full likelihood, RE(FL) and penalized quasi-likelihood, RE(PQL). The results showed that on average, OLR yielded the smallest standard errors, GEE the second smallest, and random effects model the largest. Among the three methods accounting for correlated observations, GEE tended to yield the smallest standard errors. Their findings supported the theory that the random effects models provide the largest standard error of regression coefficients.

Bellamy et al conducted a simulation study, comparing the type I error rate and power of OLR, GEE_{ind} , GEE_{exc} and RE(PQL) methods with an intraclass correlation equal to 0.10 with various settings of the number of clusters and number of subjects in each cluster. It was shown that type I error rate of OLR was high, roughly 20-30%. The RE(PQL) method had a consistently smaller type I error rate than GEE_{ind} and GEE_{exc} . The two GEE approaches had almost the same power. The power of the GEE methods is higher than those of RE(PQL) methods. As the number of clusters increased, the differences of power between RE(PQL) and the two GEEs decreased (Bellamy et al., 2000). Their findings were supported by a simulation study by Austin. His study examined various statistical methods including GEE and RE(PQL) and allowed the intraclass correlation to vary (0.01, 0.06, 0.11 and 0.66). The results indicated that GEE provided a higher type I error rate than RE(PQL), but also provided greater power than RE(PQL). In contrast with Bellamy et al's study, the power of GEE_{ind} and GEE_{exc} were not negligible in which the scenario was set as a class correlation was 0.11 (Austin, 2007).

Empirical studies

Hu et al investigated parameter estimates in the three methods based on longitudinal smoking status data. The standard errors for time-invariant covariates such as sex, race and treatment group (linear trend) are smaller in OLR model, whereas the standard error for time-varying covariates such as time and the interaction terms are larger in the GEE_{exc} models. The standard errors from the random effects model are larger than those from the GEE_{exc} , although the test statistics are relatively close. Their results confirmed theoretical literature.

Another three studies (Ananth, Platt, & Savitz, 2005; Carriere & Bouyer, 2002; Kuchibhatla & Fillenbaum, 2003) evaluating the three methods using real data also showed results supporting the theory. They reported the same conclusion that random effects model provides the largest standard errors regardless of the type of covariates. OLR overestimates standard errors of time-varying covariates, and underestimates the standard errors of time-invariant covariates. One of these studies (Ananth, Platt, & Savitz, 2005) revealed that OLR provided standard errors of time-invariant covariates that could range from 7-71% smaller than those from the GEE model. However, the 95% confidence intervals of the two methods are similar. The authors discussed that this is because of the large sample size (285,226 twins). In addition, the study of Kuchibhatla showed that even though the conclusion of covariate effects are the same, OLR model provided a larger p-value than those of both GEE and random effects models.

The summary of literature review of statistical properties of the three statistical methods based on theoretical, simulation and empirical studies is shown in table 1, 2 and 3 respectively

Table 1 Summary of the statistical properties of ordinary logistic regression, generalized estimating equations and random effects models based on theories

	Ordinary logistic regression (OLR)	Generalized estimating equations (GEE)	Random effects models (RE)
General form of the model	$\Pr(Y_{it}) = f(X_{it} \beta_{OLR})$	$\Pr(Y_{it}) = f(X_{it} \beta_{GEE})$	$\Pr(Y_{it}) = f(X_{it} \beta_{RE} + \mu_i)$
Statistical approach	Model the population averaged expectation of the dependent variable as function of the covariates but not accounting for non-independence across observations of time	Model the marginal (population-averaged) expectation of the dependent variable as function of the covariates	Model the probability distribution of the dependent variable as function of the covariates and a parameter specific to each subject
Concept of analyzing data	Model regression of Y on X by assuming any observation is independent	Model regression of Y on X and the within-subjects dependence separately	Model regression of Y on X and the within-subjects dependence simultaneously
Term used to handle dependency between observations	None	Working correlation structure	Unit-specific effect (μ_i)
Regression coefficient estimates ($\hat{\beta}$)	Less precise than $\hat{\beta}_{GEE}$ and $\hat{\beta}_{RE}$	$\hat{\beta}_{GEE} < \hat{\beta}_{RE}$	$\hat{\beta}_{RE} > \hat{\beta}_{GEE}$
Variance estimate or standard error of $\hat{\beta}$	Incorrect standard error of $\hat{\beta}$ Overestimate for time-varying covariate Underestimate for time-invariant covariate	Correct standard error of $\hat{\beta}$	Correct standard error of $\hat{\beta}$ $SE(\hat{\beta}_{RE})$ is generally larger than $SE(\hat{\beta}_{GEE})$ and $SE(\hat{\beta}_{OLR})$
Regression coefficient in the model represents	The average effect of a one-unit change in X_{it} on $\Pr(Y_{it})$	The average effect across the entire population of a one-unit change in X_{it} on $\Pr(Y_{it})$	The effect of a change in X_{it} on $\Pr(Y_{it})$ for the same individual i
The method is used for	Investigating the effect of covariates across population when each observation is independent	Making comparisons across groups or subpopulations	Evaluating the effect of change in individuals' responses across time within a particular observation

Table 2 Summary of the statistical properties of ordinary logistic regression, generalized estimating equations and random effects models based on simulation studies

Authors (year)	Parameter settings	Repetitions per simulation	Statistical methods	Statistical properties under investigation	Findings
Heo et al (2005)	no. of cluster = 20, 100, 250 no. of obs./cluster = 8, 20, 100 $\beta = 0, 0.3, 0.5$ ICC = 0, 0.05, 0.10, 0.30	1,000	FELR or OLR GEE RE (PQL) RE(FL)	Type I error Statistical power Bias of estimates SE	<p><i>ICC $\neq 0$ condition</i></p> <ul style="list-style-type: none"> - OLR yielded the largest type I error, GEE the second highest and RE(FL) the smallest - For all methods, the larger no. of cluster provides the higher power - The power of each method tended to decrease when ICC increases regardless of different settings of β, no. of cluster and no. of obs./cluster - OLR and GEE estimates are more biased toward underestimate than RE(PQL) and RE(FL) estimates. - OLR yielded smallest SE <p><i>ICC = 0 condition</i></p> <ul style="list-style-type: none"> - GEE provided largest type I error - OLR provided smallest SE <p><i>Conclusion</i></p> <ul style="list-style-type: none"> - The performance of OLR was very sensitive to size of ICC and should be avoid when ICC $\neq 0$ - regardless of the methods, the bias of coefficients tended to increase with size of the coefficients and ICC. - RE(FL) approach are more preferable, even if within-cluster is close to zero - RE(PQL) approach performs well when no. of obs./cluster is large

Table 2 Summary of the statistical properties of ordinary logistic regression, generalized estimating equations and random effects models based on simulation studies (Cont.)

Bellamy et al (2000)	No. of cluster = 10, 20, 30, 50 No. of subj./cluster = 10, 100 ICC = 0.10 Response rate of control = 0.23 Response rate in intervention arm = 0.09, 0.13, 0.18, 0.23, 0.28	500	OLR GEE _{ind} GEE _{exc} RE(PQL)	Type I error Statistical power	<ul style="list-style-type: none"> - Type I error rate of OLR was high, roughly 20-30% - RE(PQL) produced a consistently smaller type I error than GEE_{ind} and GEE_{exc} - Power of GEE_{ind} \approx power of GEE_{exc} - As the number of clusters increased, the differences of power between RE(PQL) and two GEE methods diminished. - When no. of cluster was small (10 or 20), the RE(PQL) had moderately lower power than GEE_{exc}
Austin (2007)	no. of cluster per arm = 13,30 no. of subj./cluster = 7, 39 ICC = 0.01, 0.06, 0.66 Response rate of control = 0.5 Response rate in intervention arm = -0.35 to 0.35 in increments of 0.05	1,000	t- test Wilcoxon rank sum test Permutation test Adjust chi-square test	Type I error Statistical power	<p><i>Equal number of subjects per cluster</i></p> <ul style="list-style-type: none"> - GEE had the greater power than other five methods for any situations in combination of ICC, no. of cluster/arm and no. of subj./cluster, however, the differences were small - GEE produced a higher type I error rate than RE(PQL) for any situations in combination of ICC, no. of cluster/arm and no. of subj./cluster <p><i>Unequal number of subjects per cluster</i></p> <ul style="list-style-type: none"> - Same results as equal number of subjects per cluster.

FELR = Fixed- effect logistic regression or ordinary logistic regression, OLR = Ordinary logistic regression, GEE =Generalized estimating equations, RE(PQL) = Random effects models based on penalized quasi-likelihood method, RE(FL) or RE = Random effects models based on full likelihood method, ICC = Intraclass correlation

Table 3 Summary of the statistical properties of ordinary logistic regression, generalized estimating equations and random effects models based on empirical studies

Authors (year)	Population	Design/Method of study	Statistical methods	Covariates/Outcome	Findings
Zorns (2001)	1,734 voter of House Judiciary Committee	Cross-sectional study with correlated outcome Vote out four articles of impeachment against President Clinton in the same day	GEE _{ind} GEE _{exc} GEE _{uns} RE	<i>Time-invariant cov.</i> - Clinton's 1996 vote percentage - Republican ember - D-NOMINATE Score <i>Outcome</i> Impeachment	- RE estimators are uniformly larger than those from all GEE methods - $\hat{\beta}_{GEE(exc)} = \hat{\beta}_{GEE(ind)}$ (in the three decimal places)
Neuhaus et al (1991)	940 samples of fluid from both breast of women at University of California	Cross-sectional study with correlated outcome Fluid from both breast of the same woman was investigated for Dysplasia	GEE _{exc} RE	<i>Cluster-constant cov.</i> Age Age at menarche Full term birth <i>Outcome</i> Dysplasia	- $\hat{\beta}_{RE} > \hat{\beta}_{GEE(exc)}$ in all covariates - $\hat{\beta}_{GEE(exc)}$ are closer to zero than $\hat{\beta}_{RE}$
Hu et al (1998)	1,607 Adolescents from Midwestern Prevention Project	Longitudinal study 7 time points follow up for smoking	OLR GEE _{exc} RE Stratified analysis (Mantel-Haenzel method) Conditional logistic	<i>Time-invariant cov.</i> Sex Race Grade Baseline smoking Treatment group <i>Time varying cov.</i> Time Time *group <i>Outcome</i> Smoking	<i>When analyzing two time points (baseline and one year)</i> - Odds ratios of treatment effect and time effect from logistic models are nearly identical to those from stratified analyses - $\hat{\beta}_{OLR} = \hat{\beta}_{GEE(exc)}$ for time effect - Odds ratio of time effect from RE models and conditional logistic are identical <i>When analyzing seven time points</i> - $\hat{\beta}_{RE} > \hat{\beta}_{GEE(exc)} \geq \hat{\beta}_{OLR}$ - $SE(\hat{\beta}_{OLR})$ is overestimated for time-varying covariates and underestimated for time-invariant covariates

Table 3 Summary of the statistical properties of ordinary logistic regression, generalized estimating equations and random effects models based on empirical studies (Cont.)

Authors (year)	Population	Design/Method of study	Statistical methods	Covariates/outcome	Findings
					<ul style="list-style-type: none"> - $SE(\hat{\beta}_{RE}) > SE(\hat{\beta}_{GEE(exc)}) > SE(\hat{\beta}_{OLR})$ - All $\hat{\beta}$ from GEE and RE showed relatively close results in making inference of significance effect of covariates
Carriere & Bouyer (2002)	1,548 Elderly women in Montpellier, Southern France	Longitudinal study Self-reported disability in older women assessed annually for 6 years	GEE _{exc} RE	<i>Time-invariant cov.</i> Age BMI Visual acuity Perceived health <i>Time-varying cov.</i> Hospitalized Temporarily confined <i>Outcome</i> Disability	<i>All covariates</i> <ul style="list-style-type: none"> - $\hat{\beta}_{RE} > \hat{\beta}_{GEE(exc)}$ - All $\hat{\beta}$ from GEE and RE showed the same results in making inference of significance effect of covariates <i>All covariates</i> <ul style="list-style-type: none"> - $SE(\hat{\beta}_{RE}) > SE(\hat{\beta}_{GEE(exc)})$
Kuchibhatla et al (2003)	2,231 elderly from EPESE	Longitudinal study 3 time points follow up for cognitive status	OLR GEE _{exc} RE	<i>Time-invariant cov.</i> Age Gender ADL problems IADL problems <i>Time-varying cov.</i> Time Time by ADL Time by IADL <i>Outcome</i> Cognitive impair.	<i>All covariates</i> <ul style="list-style-type: none"> - $\hat{\beta}_{RE} > \hat{\beta}_{GEE(exc)}, \hat{\beta}_{RE} > \hat{\beta}_{OIR}$ - $\hat{\beta}_{OIR}$ could be either bigger or smaller than $\hat{\beta}_{GEE(exc)}$ <i>Time-invariant covariates</i> <ul style="list-style-type: none"> - $SE(\hat{\beta}_{OLR}) < SE(\hat{\beta}_{GEE(exc)})$ <i>Time-varying covariates</i> <ul style="list-style-type: none"> - $\hat{\beta}_{OIR} > \hat{\beta}_{GEE(exc)}$

Table 3 Summary of the statistical properties of ordinary logistic regression, generalized estimating equations and random effects models based on empirical studies (Cont.)

Author (year)	Population	Design/Method of study	Statistical methods under comparison	Covariates/outcome	Findings
Ananth et al (2005)	142,613 Twins	Cohort study with correlated response Twin pregnancies were evaluated for the risk of perinatal death	OLR GEE _{ind} GEE _{exc} RE	<p><i>Cluster-constant cov.</i> Gestation age Race Primigravida Multigravida with prior PTD Placental abruption Intrapartum fevemniosr Hydra</p> <p><i>Cluster-varying cov.</i> Birth weight Second twin Smaller twin Male</p> <p><i>Outcome</i> Perinatal death</p>	<p><i>All covariates</i> - $\hat{\beta}_{OLR} = \hat{\beta}_{GEE(ind)}$ - $\hat{\beta}_{RE} > \hat{\beta}_{GEE(exc)}$, $\hat{\beta}_{RE} > \hat{\beta}_{OLR}$ - $\hat{\beta}_{OLR}$ could be either bigger or smaller than $\hat{\beta}_{GEE(exc)}$</p> <p><i>Cluster-constant covariates</i> - $SE(\hat{\beta}_{OLR}) > SE(\hat{\beta}_{GEE(exc)})$</p> <p>- $SE(\hat{\beta}_{OLR})$ were underestimate by a factor 7-70% when compare with $SE(\hat{\beta}_{GEE(exc)})$, however the 95%CI are similar</p> <p><i>Cluster-varying covariates</i> - $SE(\hat{\beta}_{OLR})$ may be either over- or underestimated, depending on the correlation and distribution of covariates</p>

OLR = Ordinary logistic regression, GEE_{ind} = Generalized estimating equations with independent correlation structure, GEE_{exc} = Generalized estimating equations with exchangeable correlation structure, GEE_{uns} = Generalized estimating equations with unstructured correlation structure, RE = Random effects models, PTD = Preterm delivery, EPESE = Established Populations for Epidemiologic Studies of the Elderly

2.7 Review of Speech and Language Development

Language and speech are the expressions of human communication through which ideas, information, emotion and beliefs can be shared. Speech is the motor act of communicating by articulating verbal expression, whereas language is the knowledge of a symbol system used for interpersonal communication (Blum & Baron, 1997). The most intensive period of speech and language development for humans is during the first three years of life, a period when the brain is developing and maturing.

Speech refers to the actual process of making sounds, using such organs and structures as the lungs, vocal cords, mouth, tongue, teeth, etc. Speech delay refers to a delay in the development or use of the mechanisms that produce speech. In general, a child is considered to have speech delay if the child's speech development is significantly behind the norm for children of the same age. A child with speech delay has the same speech development as a child of younger chronologic age; the speech delayed child's skills are developed in a normal sequence but at a slower than normal rate (Ansel, Landa, & Stark-Selz, 1994).

A speech delay can be caused by problems with the output of speech (anatomical problems with the vocal cords, etc), the input of speech (hearing loss), or the processing of speech (mental retardation and developmental language disorders). Therefore, the two main types of speech delay are *expressive delay*, which is the inability to generate speech or to formulate vocabulary, and *receptive delay*, the inability to decode or understand the speech of others. Children can have a delay with a mix of types (mixed expressive/receptive delay).

Prevalence rates of speech and language delay vary across studies. This is because of differences in diagnostic criteria, unreliability of unconfirmed parental observations and methodological problems in sampling and data retrieval (Leung & Kao, 1999). According to recent literature reviews on speech and language development problems in preschool-aged children (2 to 4.5 years old), the prevalence of combined speech and language delay ranges from 5% to 15% (Burden, Scott, Forge, & Goodyer, 1996; Feldman, 2005); the prevalence of language delay is approximately 2.3% to 19% (Law, Boyle, Harris, Harkness, & Nye, 1998; Rescorla, Hadicke-Wiley, & Escarce, 1993); the prevalence of speech delay is between 3 and 14% (Campbell, 2000; Leung & Kao, 1999).

Preschool children with speech and language delays are at high risk for persisting language impairment and academic learning difficulties (Bishop, Venables, & Wang, 2004; Lewis, Freebairn, & Taylor, 2000), as well as with behavioral/social-emotional problems (Cantwell & Baker, 1991; Redmond & Rice, 1998).

Slow speech and language development in preschoolers is the greatest concern among parents but expressive speech delay or late talking is often not diagnosed until about 3 years of age. Parents may not be aware of what composes normal development, and most pediatricians may be hesitant to diagnose the child as delayed without thorough testing due to a wide variation in normal language development (Eapen, Zoubeidi, & Yunis, 2004).

A number of screening procedures and methods have been proposed and most of these screening programs are conducted and used in North America and Europe. Very few studies have been undertaken in Asia. In Thailand, there have not been any studies

using screening tests to assess child's language development. Generally, milestones for speech and language development in children are qualified to be used for evaluating speech and language development.

2.7.1 Normal Speech and Language Development Milestones

Children vary in their development of speech and language. There is a natural progression or timetable for mastery of these skills. The milestones are identifiable skills that serve as a guide to normal development. Typically, speech development progresses through the stages of cooing, babbling, echolalia, jargon, words and word combination, and sentence formation (Leung & Kao, 1999). A child may be considered at risk for having speech and language delay if he/she has not achieved a milestone by the age indicated. The normal speech/language development milestones are shown in Table 4 (Schwartz, 1990).

Table 4 Normal speech and language development milestones

Age	Achievement
1 to 6 months	Coos in response to voice
6 to 9 months	Babbling
10 to 11 months	Imitation of sounds; says "mama/dada" without meaning
12 months	Says "mama/dada" with meaning; often imitates two-and three-syllable words
13 to 15 months	Vocabulary of four to seven words in addition to jargon; <20% of speech understood by strangers
16 to 18 months	Vocabulary of 10 words; some echolalia and extensive jargon; 20% to 50% of speech understood by strangers
19 to 21 months	Vocabulary of 20 words; 50% of speech understood by strangers
22 to 24 months	Vocabulary >50 words; two-words phrases; dropping out of jargon; 60% to 70% of speech understood by strangers
2 to 2½ years	Vocabulary of 400 words, including names; two- to three-word phrases; use of pronouns; diminishing echolalia; 75% of speech understood by strangers
2½ to 3 years	Use of plurals and past tense; knows age and sex; counts three objects correctly; three to five words per sentence; 80% to 90% of speech understood by strangers
3 to 4 years	Three to six words per sentence; asks questions, converses, relates experiences, tells stories; almost all speech
4 to 5 years	Six to eight words per sentence; names four colors; counts 10 pennies correctly

2.7.2 Common Causes of Speech and Language Disorders

Speech and language delays may be the manifestation of numerous disorders.

Causes of speech and language problems include the following:

Mental retardation is a developmental disability that is marked by lower-than-normal intelligence and limited daily living skills. A mentally retarded child demonstrates global language delay and also has delayed auditory comprehension and delayed use of gestures. Mental retardation accounts for more than 50 percent of language delays.

Hearing impairment is one of the most common causes of language delay. Any child who does not hear speech in a clear and consistent manner will have language delay. Even a minor hearing impairment can significantly affect language development. Hearing loss may be conductive or sensorineural. Conductive loss is commonly caused by otitis media with effusion whereas sensorineural hearing loss may result from intrauterine infection, kernicterus and bacterial meningitis.

Autism is a neurologically based developmental disorder; onset occurs before the child reaches the age of 36 months. Autism is characterized by delayed and deviant language development, failure to develop the ability to relate to others and ritualistic and compulsive behavior, including stereotyped repetitive motor activity.

Maturation delay or developmental language delay accounts for a considerable percentage of late talkers. In this condition, a delay occurs in the maturation of the central neurological process required to produce speech.

Cerebral palsy is a disorder of movement and posture caused by an early permanent or non-permanent cerebral lesion. Speech delay occurs most often in those with an athetoid type of cerebral palsy.

2.8 Risk Factors Related to Speech and Language Development

Sixteen studies that have assessed risk factors include three case-control studies, four cohort studies, four cross-sectional studies and three longitudinal studies. They were reviewed for the purpose of searching for potential risk factors influencing speech and language development in children. The significant risk factors (variables) from the literature review will be considered as candidate variables for data analysis using three statistical methods; OLS, GEE and random effects models. The findings of the literature review of biologic and environmental factors related to speech and language delays are described subsequently.

Family History of Language Developmental Delays

Family history is defined as family members who were late to talk or had language disorders, speech problems, or learning problems. Family history is one of the most consistent risk factors of language disorders. A cross-sectional study in children aged ½ - 5 years found that the language scores of children with a family history of language disorders were at or below the 10th percentile of those of children without family history of language disorders (Tomblin, Hardy, & Hein, 1991). This result was supported by many case-control studies, which indicated that positive history of speech and language problem significantly increases the risk of speech and language delay (Campbell, 2000; Choudhury & Benasich, 2003; Fox, Dodd, & Howard, 2002; Tomblin,

Smith, & Zhang, 1997). Among these studies, the biggest effect was reported to be about 4.38 times as likely for a child having positive family history of speech and language difficulties (Fox, Dodd, & Howard, 2002). The significant relationship between family history of language delay and developmental language delay (DLD) was also found in a survey study using screening procedures in detecting DLD in children aged 3 years (Eapen, Zoubeidi, & Yunis, 2004). Most recently, a study also suggested that having a family history of speech and language difficulties was more likely to lower speech scores (Reilly et al., 2007).

Low Birth Weight and Premature Birth

A few studies have shown that low birth weight has a negative impact upon a child's later language development. Stanton-Chapman et al studied a large cohort of children at ages 6-7. They found that children with very low birth weight had a higher risk of language impairment than those with moderate-low birth weight and normal birth weight (Stanton-Chapman, Chapman, Bainbridge, & Scott, 2002). Results from this study are consistent with previous studies that have identified low birth weight as being associated with language disorders (Weindrach, Jennen-Steinmetz, Laucht, Esser, & Schmidt, 1998; Yliherva, Olsn, Maki-Torkko, Koironen, & Jarvelin, 2001). However, in Tomblin et al study, they did not find significant association between low birth weight (< 2,500 gm) and SLI (Tomblin, Smith, & Zhang, 1997). Cusson conducted a longitudinal study to evaluate language development of preterm infants, who were below 2,000 gm and less than 36 weeks gestation at birth. He found that by 26 months age, infant development was within the normal range but expressive and receptive language was delayed an average of 3 to 5 months (Cusson, 2003). This result is consistent with a

previously longitudinal study which found that 5-year-old children who were born preterm, especially those born at 32 weeks gestation or earlier, were more likely to have poorer language comprehension, language production, and speech ability than their pairwise-matched controls and than children born full term. A strong relationship between prematurity and specific language impairment (SLI) were also reported in the study by Weindrich et al. The odd ratios were 2.2 and 2.8 for expressive and receptive SLI respectively (Weindrich, Jennen-Steinmetz, Laucht, Esser, & Schmidt, 1998). However, a study of population-based investigation of birth risk factors by Stanton-Chapman and colleagues failed to find a significant association between prematurity (<37 weeks gestation) and SLI (Stanton-Chapman, Chapman, Bainbridge, & Scott, 2002).

Gender

One of the most extensively studied factors influences a child's language development is the child's gender. Morisset et al investigated language development in preschool children; the results showed that girls appear to have a slight advantage in vocabulary acquisition during early childhood (Morisset, Barnard, & Booth, 1995). Many studies came to the same conclusion that male gender was a significant factor for language development and also reported a consistently prevalence of boy-to-girl ratio, which ranges from 1.2 to 3.1 (Campbell, 2000; Choudhury & Benasich, 2003; Reilly et al., 2007; Shriberg, Tomblin, & McSweeny, 1999).

Breastfeeding

Breastfeeding is widely regarded as beneficial for child health including growth and development. Tomblin and colleagues found a significant protective effect of breastfeeding on SLI. Children who breast fed longer had a lower risk for SLI, compared to children who never breast fed or had a short duration of breast feeding (Tomblin, Smith, & Zhang, 1997). Another study related to mother's concern about her child's language development, confirmed this result by showing that breast feeding may protect against speech delay in children. Fewer concerns about language delay were apparent for children breast fed ≥ 3 months, and mother's concerns generally reduced as breast feeding continued ≥ 9 months (Dee, Ruowei, Li-Ching, & Grummer-Strawn, 2007).

Apgar Scores

The Apgar score is an indicator of the infant's postnatal condition. Apgar scores assessment has been accepted as a basis for prediction of later mental and motor development. Andrew et al reported that Apgar scores less than 8 was a significant predictor for future SLI (Andrew, Goldberg, Wellen, Pittman, & Struening, 1995). However, a recent study showed that low Apgar scores (<6) at 5 min were associated with about double the risk for SLI at ages 6-7 years, compared to children with Apgar scores of 7 or higher (Stanton-Chapman, Chapman, Bainbridge, & Scott, 2002).

Birth Order

Birth order was reported less consistently as a risk factor of speech and language disorders. One study failed to find the effect of birth order on language development disorders after controlling for family size and socioeconomic status (Tomblin, 1990).

However, some recent studies reported that birth order was a significant factor. Their findings were that the higher order born child was likely to have inferior language ability and having a higher risk of SLI than the first born child. Those children who were third born or later have also been associated with later language impairment (Horwitz et al., 2003; Stanton-Chapman, Chapman, Bainbridge, & Scott, 2002; Tomblin, Hardy, & Hein, 1991).

Parental Education

Three studies demonstrated that parental education is associated with children's language development. Tomblin et al revealed that either maternal or paternal education influences their children's language development in the same manner. This means that, the lower level of education among parent, the higher the risk of SLI among their children (Tomblin, Smith, & Zhang, 1997). Weindrich supported the effect of this factor. He found that children from low educated mothers have a significantly higher risk of expressive language disorder than those from high educated mothers (Weindrich, Jennen-Steinmetz, Laucht, Esser, & Schmidt, 1998). Dollaghan et al showed a significant linear trend of mean scores for all language skills and increasing maternal education level. The mean scores of language of children whose mothers had not completed high school were significantly lower than children whose mothers had graduated from high school or college (Dollaghan et al., 1999). Reilly et al reported that mother's education level (≥ 13 years), were strongly associated with higher scores on speech composites (Reilly et al., 2007). Their results are consistent with previous study, finding that children born to a low educated mother (< 12 years) were twice as likely to have language

impairment than those who were from highly educated mothers (Stanton-Chapman, Chapman, Bainbridge, & Scott, 2002).

Maternal Age

Only one study evaluated the relationship between maternal age and language development outcome, for instance SLI. A cross-sectional study by Stanton-Chapman et al revealed that children of young mothers (age<18 years) had a slightly higher risk for SLI than children whose mothers were aged ≥ 18 years (Stanton-Chapman, Chapman, Bainbridge, & Scott, 2002).

Socioeconomic Status

Socioeconomic status is one demographic factor which has an impact upon individual child development. Singer et al investigated the medical complications of prematurity and the socioeconomic status factor. The results showed that low socioeconomic status was a significant factor predicting speech and language delay (Singer et al., 2001). This result was supported by a case-control study that found that children from families having private health insurance tend to be at a lower risk of speech delay than children from families having Medicaid health insurance with OR = 1.59 (Campbell, 2000). A survey study using a screening test revealed that total monthly income of the family is also associated with language delay (Eapen, Zoubeidi, & Yunis, 2004). However, Horwitz et al's study found no significantly different risk of expressive language delay among children living in poverty or near poverty, middle class household and high class households (Horwitz et al., 2003).

In summary, the literature suggests that the factors that significantly enhance the risk of speech and language delay include family history of language disorder, gender, low birth weight and premature birth, breastfeeding, Apgar score, birth order, maternal education, maternal age and socioeconomic status. Among these factors, family history was the most consistent significantly associated risk factor. The male gender appeared as a strong significant risk factor across all studies that examined it. Low birth weight and socioeconomic status were reported less consistently. Other risk factors showed a similar pattern of significant association with speech and language delay except premature birth. A summary of the studies investigating risk factors of speech and language delay is shown in Table 5.

Table 5 Summary of studies evaluating risk factors affecting speech and language development

Authors (year)	Country	Age, month	N	Type of Study	Outcome measurement	Factors	Magnitude of effect	Confidence interval
Tomblin et al (1997)	USA	Kindergarten Children	177: 925	Case control study	SLI	Maternal history No history of speech/language problem Speech problem Mental retardation Learning disability SP, MR or LD Paternal history No history of speech/language problem Speech problem Mental retardation Learning disability SP, MR or LD Birth weight Normal birth weight Low birth weight Maternal education Complete college Incomplete college Complete high school Incomplete high school Paternal education Complete college Incomplete college Complete high school Incomplete high school Breast feeding No Yes	Reference OR = 0.8 OR = 1.5 OR = 1.6 OR = 1.3 Reference OR = 1.9 OR = 3.9 OR = 1.8 OR = 2.1 Reference OR = 1.7 Reference OR = 1.7 OR = 2.3 OR = 3.5 Reference OR = 1.8 OR = 2.4 OR = 3.2 Reference OR = 0.5	 0.43, 1.5 0.60, 3.8 0.87, 2.0 0.81, 2.0 1.1, 3.3 1.9, 8.1 1.0, 3.3 1.3, 3.1 0.8, 3.8 1.1, 2.8 1.6, 3.9 1.8, 5.5 1.1, 3.0 1.6, 3.8 1.8, 5.5 0.4, 0.7

Table 5 Summary of studies evaluating risk factors affecting speech and language development (Cont.)

Authors (year)	Country	Age, month	N	Type of Study	Outcome measurement	Factors	Magnitude of effect	Confidence interval
Campbell et al (2003)	USA	3 yrs	100	Case control study	Speech delay	Positive family history Male sex Low maternal education Medicaid health insurance	OR = 1.67 OR = 2.19 OR = 2.58 OR = 1.59	1.06, 2.62 1.38, 3.47 1.06, 2.62 1.02, 2.49
Choudhury et al (2003)	USA	3 yrs	136	Cohort study	Language development	Family history Negative family history Positive family history Gender Female Male	Reference OR = 5.48 Reference OR = 1.91	1.33, 26.30
Fox et al (2002)	England	2.7 – 7.2 yrs	65: 84	Case control study	Speech disorder	Family history Negative family history Positive family history	Reference OR = 6.81	2.04, 18.66
Reily et al (2007)	Australia	8 -24 ms.	1,720	Longitudinal study	Speech and language score CSBS CDI	Family history of speech/language difficulties Female sex Maternal education ≤ 12 yrs. 13 yrs University degree Postgraduate degree	OR = 1.58 OR = 0.86 Reference OR = 0.62 OR = 0.67 OR = 0.67	1.18, 2.11 0.66, 1.12 0.44, 0.87 0.45, 0.99 0.99, 1.05
Stanton-Chapman et al (2002)	USA	6-7 yrs	5,862: 201,830	Cross-sectional study	Specific language impairment	Birth weight Normal birth weight VLBW(1,500 g) MLBW(1,500-2499 g) Gestation age Full-term (37-42 wks) Preterm(<37 wks) Apgar scores 7-10 4-6 0-3	Reference OR = 2.2 OR = 1.4 Reference OR = 0.9 Reference OR = 1.3 OR = 2.0	1.8, 2.8 1.2, 1.4 0.8, 1.0 1.0, 1.7 1.3, 3.3

Table 5 Summary of studies evaluating risk factors affecting speech and language development (Cont)

Authors (year)	Country	Age, month	N	Type of Study	Outcome measurement	Factors	Magnitude of effect	Confidence interval
						Birth order First or second Third or fourth Fifth or more Maternal education >12 yrs = 12 yrs < 12 yrs Maternal age 18-35 yrs < 18 yrs	Reference OR = 1.3 OR = 1.4 OR = 1.4 Reference OR = 1.6 OR = 1.3 Reference OR = 1.1	1.2, 1.4 1.3, 1.6 1.3, 1.6 1.5, 1.7 1.2, 1.4 1.0, 1.2
Ylihera et al (2001)	Finland	8 yrs	9322	Cohort study	Speech production Speech perception	Birth weight ≥ 2500 g 1500-2499 g < 1500 g Birth weight ≥ 2500 g 1500-2499 g < 1500 g	Reference OR = 1.5 OR = 1.1 Reference OR = 2.0 OR = 1.5	1.1, 2.6 0.4, 2.6 0.2, 5.3 1.2, 3.2
Cusson (2002)	USA	7-26 ms	43	Longitudinal study	Expressive and receptive language delay	Preterm infants birth weight < 2,000 g	Delayed an average of 3-5 months	N/A
Weindrich et al (1998)	Germany	2- 4 5 yrs	324	Longitudinal study	Expressive language	Premature birth Full term birth Premature birth Birth weight > 2500 g < 2500 g Parental education High educated Low educated	Reference RR = 2.2 Reference RR = 4.4 Reference RR = 2.0	N/A N/A N/A

Table 5 Summary of studies evaluating risk factors affecting speech and language development (Cont.)

Authors (year)	Country	Age, month	N	Type of Study	Outcome measurement	Factors	Magnitude of effect	Confidence interval
					Receptive language	Premature birth Full term birth Premature birth Birth weight >2500 g ≤ 2500 g Parental education High educated Low educated	Reference RR = 2.8 Reference RR = 0.8 Reference RR = 2.5	N/A N/A N/A
Dee et al (2007)	USA	10-17 ms	22,399	Cross sectional study	Parental concern about their child's language development Expressive language Receptive language	Breastfeeding Yes No Breastfeeding Yes No	Reference OR = 0.78 Reference OR = 0.70	0.67, 0.91 0.6, 0.81
Horwitz et al (2003)	USA	12-39 ms	870	Cross-sectional study	Expressive language delay (MCDI)	First born No Yes Maternal education College degree < College degree SES Borderline poverty Poverty	Reference OR = 0.48 Reference OR = 1.26 Reference OR = 1.33	0.29, 0.81 0.74, 2.03 0.78, 2.27

Table 5 Summary of studies evaluating risk factors affecting speech and language development (Cont.)

Authors (year)	Country	Age, month	N	Type of Study	Outcome measurement	Factors	Magnitude of effect	Confidence interval
Dollaghan et al (1999)	USA	3 yrs	240	Cross-sectional study	Speech and language score MLUm NDW TNW PPVT-R PCC	<i>Maternal education</i> < High school, High school, College were tested for linear trend	Linear trend F = 22.80 F = 24.45 F = 8.41 F = 3.43 F = 74.64	P-value <.0001 <.0001 <.01 0.65 <.0001
Singer et al (2001)	USA	3 yrs	163	Cohort study	Language delay	SES	Coefficient $\beta = -0.25$	P-value .001
Eapen et al (2004)	United Arab Emirates	3 yrs	694	Cross-sectional study	Developmental language delay DDST	Family history of DLD	OR = 2.54 OR = 0.74	1.37, 4.70 0.60, 0.93

SLI = Specific Language Impairment, Communication and Symbolic Behavior Scale (CSBS), MacArthur-Bates Communicative Development Inventory (MCDI), OR = Odds Ratio, SES = Socioeconomic status, MLUm = Mean length of Utterance in morphemes, NDW = Number of Difference Words, TNW = Total Number of Words, PCC = Percentage of Consonants Correct, PPVT-R = Peabody Picture Vocabulary Test-Revised standard score, DDST = Denver Developmental Screening Test, DLD = Developmental Language Delay, N/A = Not available

CHAPTER III

METHODOLOGY

The primary goal of this study is to compare parameter estimates and their standard errors generated by three different estimation techniques: the ordinary logistic regression model, the marginal logistic regression model fit using the GEE and the random effects logistic regression model. A second goal is to assess risk factors of speech delay using a longitudinal study with a binary outcome. The details of the methods used to achieve these goals are described below:

3.1 Data set

The data used in this study are taken from the Prospective Cohort Study of Thai Children (PCTC), a longitudinal study on comprehensive multidisciplinary of the biological, psychological and moral development in Thai children. The population base for this cohort study included all pregnant women in community-based studies from four regions and one hospital-based study from Bangkok, the capital of Thailand. The criteria for eligibility were gestational age 28th to 38th week and willingness to participate. This is a cohort with a two year range of entry. The date of entry into the birth cohort study was the day of the child's birth, between October 15, 2000 and September 14, 2002. There were 4,245 children enrolled in the study. The birth cohorts will be observed and followed until 24 years of age. Data on pregnancy outcomes and child development, including secondary data regarding community and demographic information were collected from all subjects. Most assessments were carried out at home, and some were

administered at the hospital. The PCTC study is a project that assesses child development in many areas, including mental, cognitive, language, emotional and moral development; however, this study focuses only on language development regarding speech delay with three waves of assessments at the age of 12 months, 18 months and 24 months.

Two data sets were derived from the full data set (N = 4,245). The complete data set consists of children who have no missing value on all predictor variables as well as outcome variables across three times of measurement (N =1,823). The complete at baseline data set consists of children who have complete data on all predictor variable and outcome of SSD at three time points, and children who have one missing value on either outcome of SSD at 18 or 24 months (N = 2,925).

3.2 Outcome and Covariates

The binary outcome for this study is suspected speech delay. The indicator of suspected speech delay is positive if the child is not able to achieve the language development milestone appropriate for his or her age. The criteria for each age are listed below:

At 12 months: no single word can be pronounced

At 18 months: vocabulary less than 10 words

At 24 months: vocabulary less than 50 words

This study uses the term “*suspected speech delay*”(SSD) rather than “*speech delay*” because there is no universal tool or test to definitely diagnose speech delay in children and the definition of speech delay in Thailand has not been established. Therefore, the SSD status of children in the PCTC study using criteria above are

considered as a screening assessment which is an initial indication of speech delay in Thai children. The outcome was observed by parents and caregivers. They were trained to assess the language ability using a special monthly calendar for developmental record, developed by experts in the research team. Other information on risk factors were obtained from hospital records and interviews conducted by research assistants.

The covariates considered for the data analyses are the risk factors for suspected speech delay suggested by the literature. They are family history of language disorder, birth order, low birth weight, breast feeding, Apgar scores, parental education, maternal age and socioeconomic status. It is noted that family history is not included since this variable is not available in the dataset. The description and coding for the variables in the analysis are shown in Appendix A (Table A1)

3.3 Statistical Models

This study considers logistic regression models based on the population-averaged and the subject-specific approaches. All different models derived from the two methods were applied in both data sets for evaluating parameter estimates and identifying risk factors for SSD

Population-averaged or marginal logistic regression models

The marginal model will be fit using three different methods of estimation.

1. The ordinary logistic regression in which the time dependency is ignored. This model is fit using PROC LOGISTIC.

2. The GEE logistic regression based on independent correlation structure which assume all repeated observations are not correlated. The SAS procedure for fitting the model is PROC GENMOD with subcommand *independent*.
3. The GEE logistic regression based on exchangeable correlation structure in which the correlations across repeated observations are assumed to be equal. The command PROC GENMOD with subcommand *exchangeable* is use for this method.

Subject-specific logistic regression models

The subject-specific model will be fit using random effects models, where the individual-specific effects are assumed to be distributed as $N(0, \tau^2)$. For this model, the SAS command based on the Gauss-Hermite quadrature integration method, PROC NLMIXED will be used. Under the random effects procedure, two random effects models will be performed: random intercept model and random slope for the time covariates model.

3.4 Statistical Analysis

The data analysis chapter of this study contains two sections. Each section serves to achieve the purpose of comparing the regression parameter and standard errors from all three techniques, and assessing risk factors of speech delay. All analyses will be conducted using the SAS statistical software.

3.4.1 Comparing Regression Parameters and Standard Errors

This strategy considers analyzing binary repeated observations and focuses on comparing parameter estimates and standard errors obtained from three statistical approaches: OLR, GEE, and random effects model. The first method does not take correlations into account in the inferential process whereas the last two methods do. The ordinary logistic regression is included for providing an incorrectly specified point of reference to which results from the other two methods can be compared with respect to the assessment of the sensitivity of ignoring dependence observations. To compare models, all three logistic regression models will be fit to the data. The criteria for evaluating the three techniques are the parameter estimates and standard errors. Also, statistical tests for assessing significance such as the Wald test will be evaluated across models. In the model fitting process, all explanatory variables including time will be included in the model regardless of whether or not they are significantly associated with the outcome. This is because the purpose of this study is to compare the magnitude of the regression coefficients and their standard error of all covariates among the three methods while making the models as similar as possible.

3.4.2 Investigating Risk Factors

This section will investigate the association between risk factors and suspected speech delay. The ordinary logistic regression model, the GEE logistic model with exchangeable correlation and the random effects model with and without random slopes will be used to test the individual effects of each explanatory variable on the risk of having speech delay, while controlling for other explanatory variables in the model. All

variables with coefficients, whose p-value are less than or equal to 0.25 in the bivariate analyses will be included in the development of the model. Explanatory variables that do not contribute significantly to the models will be deleted individually based on the likelihood-ratio test. The process of testing the contribution of the explanatory variables continues until there is no variable that can be removed. It is noted that, the likelihood-ratio test is not appropriate for the GEE approach, since GEE are not based on full-information maximum likelihood. Therefore, when using GEE, the Wald test will be used to select a model.

3.5 Missing Data

The population-average and subject-specific models require different assumption of missing data mechanism. The population-average model using the GEE method needs the assumption of “missing completely at random” (MCAR). In contrast, subject-specific models using the random effects models requires more relaxing assumption of “missing at random”(MAR) (Little & Rubin, 2002). Since the main purpose of the current study is to focus the differences in the methods, exploring mechanism of missing data is beyond the scope of this work.

CHAPTER IV

RESULTS

This chapter is organized into three sections. The first section describes the demographic characteristics of the birth cohort including the results of assessing the distributions of missing data. The second section presents the results of the longitudinal analysis of risk factors related to suspected speech delay (SSD) in childhood using different statistical techniques. The ordinary logistic regression model, logistic regression using GEE with exchangeable and autoregressive (ar1) correlation structures models, a random intercept logistic regression model and a random intercept and a random slope (with linear age effect) logistic regression model are fit to data from the PCTC study. Then the parameter estimates ($\hat{\beta}$) coefficients and their standard errors ($SE(\hat{\beta})$) obtained from those five models are compared. Furthermore, the risk factors for SSD are also identified from these models. The analysis was performed on two data sets; the complete data and the complete at baseline data sets. These two data sets were derived from the full data set ($N = 4,245$). Complete data consists of children who have no missing data ($N = 1,823$). Complete at baseline data is a set of children who have no missing value, and children who have one missing value on either outcome of SSD at 18 or 24 months ($N = 2,925$). Finally, the last section discusses the comparison of results of longitudinal analysis from the complete data and complete at baseline data with respect to parameter estimates and their standard errors.

4.1 Description of Children in the PCTC Data

4.1.1 Full Data

The initial sample during the recruitment period of the longitudinal study is 4,245 children. As usual in longitudinal studies, problems of missing data are unavoidable. In the PCTC birth cohort, there were missing values on baseline characteristics and intermittent missing values on speech development assessment at months 12, 18 and 24. With three waves of assessment at age 12, 18 and 24 months, the response rates for outcome assessment were respectively: 91.5, 77.1 and 77.6%. There were 2,366 (55.7%) children who had complete observations of outcome across three time points and about 44.3% of children had at least one missing values for outcome in the three assessments. Conversely, only 46 children (1.1%) had missing value of outcome for all time points. In addition, there were 541 children who did not completely provide information on explanatory variables at baseline. Therefore, data on 1,823 children, accounting for 42.9% in a total initial sample of 4,245 children with complete records at all three time points of speech delay assessment were retained. This means that about half of the children in the birth cohort (57.1%) were omitted. Missing data occurred in all variables except gender. The missingness was concentrated on a few variables. Three variables caused a substantial amount of missing data including the outcome at 18 months, 24 months and father's education. Speech delay assessment at 18 months and 24 months were each missing in about 20% of the sample and father's education was missing for 13.1% of the children. The rest of the variables were missing for less than 10% of the sample. The number and percent of missing cases for each variable are listed in Table 6.

Table 6 Number and percent of missing data for each variables used in analysis

Variables missing	Number	Percent
Gender	0	0
Low birth weight	178	4.2
Apgar score	425	10.0
Birth order	114	2.7
Maternal age	39	0.9
Mother's education	40	0.9
Father's education	555	13.1
Breast feeding	146	3.4
Socioeconomic status	39	0.9
SSD at 12 months	360	8.5
SSD at 18 months	971	22.9
SSD at 24 months	953	22.5
Total sample = 4,245		

Since there were a substantial number of children with missing data, it is interesting to compare the data of these children to those of children with complete records. "Complete data" is a set of children with complete records on all predictor variables and the outcome (N = 1,823). The children who had at least one missing value were classified as "incomplete data" (N = 2,422). The chi-square test was used to test categorical variables and the t-test was employed to compare the difference of means for continuous variables. The results of comparisons between complete and incomplete data are presented in Table 7. Not surprisingly, many variables showed significantly differences between complete and incomplete data. Incomplete data had a slightly higher percent of children with abnormal apgar scores than complete data. The percent of illiterate and highly educated parents was double in children with incomplete data. The proportion of children who were breast fed less than 6 months was much higher for incomplete data. There was also a difference in maternal age between these two groups. According to information in Table 7, the complete and incomplete data are different, which indicates that analysis based on complete cases does not represent the total sample of the PCTC study.

Table 7 Comparison between complete and incomplete data on each covariate

Covariate	Complete cases		Incomplete cases		Statistic test	
	N	Case (%)	N	Case (%)	χ^2	p-value
Gender						
Male	1823	942 (51.7)	2422	1252 (51.7)	0.001	0.97
Female		881 (48.3)		1170 (48.3)		
Low birth wt.						
< 2,500 gms	1823	1642 (90.1)	2244	2032 (90.6)	0.31	0.58
≥ 2,500 gms		181 (9.9)		211 (9.4)		
Apgar score						
Normal (≥ 7)	1823	1816 (96.6)	1997	1977 (99.0)	5.17	0.02
Abnormal (< 7)		7 (0.4)		20 (1.0)		
Birth order						
First child	1823	1796 (98.5)	2308	2283 (98.9)	1.62	0.44
Second child		13 (0.7)		14 (0.5)		
Third child or higher		14 (0.8)		11 (0.5)		
Mathernal education						
Illiteracy	1823	69 (3.8)	2382	163 (6.9)	158.3	<.001
Primary school		1040 (57.0)		960 (40.3)		
High school		499 (27.4)		695 (29.2)		
Vocational or higher		215 (11.8)		564 (23.6)		
Paternal education						
Illiteracy	1823	43 (2.4)	1867	94 (5.0)	127.86	<.001
Primary school		1011 (55.5)		769 (41.2)		
High school		528 (29.0)		529 (28.3)		
Vocational or higher		241 (13.2)		475 (25.5)		
Breast feeding						
Less than 6 months	1823	465 (25.5)	2276	1011 (44.4)	157.13	<.001
6 months or more		1358 (74.5)		1265 (55.6)		
Socioeconomic						
Poverty	1823	628 (34.5)	2383	700 (29.4)	120.23	<.001
Borderline poverty		703 (38.6)		658 (27.6)		
Non poverty		492 (26.9)		1025 (43.0)		
Maternal age	1823	26.7 (6.20) ^a	2383	27.3 (6.33) ^a	3.08 ^b	0.002 ^b

^a mean and standard deviation ^b t-test and p-value based on t-test

4.1.2 Complete Data

Based on complete data, the baseline characteristics of children in this particular birth cohort including childbirth factors, parental factors and environmental factors are shown in Table 8. Among the 1,823 children in this sample, the proportions of gender were nearly equal; 51.7% of children were male and 48.3% were female. About 10% of

the children were reported as low birth weight (less than 2500 grams) and the average birth weight of the entire sample is 3,044.9 grams with only 0.4% of the children having Apgar scores less than 7 (indicated as abnormal). Almost all of the children (98.5%) in this cohort were firstborn. The age range for the mothers is 14 to 47, with majority of them in their twenties. The average maternal age at baseline was 26.7 with SD 6.2. About half of the fathers (55.5%) and mothers (57.0%) attained only primary school (four years in school). Moreover, approximately 2.4% and 3.8% of the fathers and mothers, respectively are illiterate. Approximately three-fourths (74.5%) of the children were breast fed more than 6 months. Regarding the socioeconomic status of the families, the majority of children (73.1%) were born to poor to borderline poor families.

Table 8 The distribution of baseline characteristics of the birth cohort based on complete data (N=1,823)

Characteristics	Number	Percent
1. Childbirth factors		
Gender		
Male	942	51.7
Female	881	48.3
Birth weight		
≥ 2,500 gms	1642	90.1
< 2,500 gms	181	9.9
Mean	3044.85	
SD	455.43	
Apgar score		
Normal (Apgar score ≥ 7)	1816	96.6
Abnormal (Apgar score < 7)	7	0.4
Birth order		
First child	1796	98.5
Second child	13	0.7
Third child or higher	14	0.8
2. Parental factors		
Maternal age		
Mean	26.69	
SD	6.20	

Table 8 The distribution of baseline characteristics of the birth cohort based on complete data (N=1,823) (continue)

Characteristics	Number	Percent
Maternal education		
Illiteracy	69	3.8
Primary school	1040	57.0
High school	499	27.4
Vocational or higher education	215	11.8
Paternal education		
Illiteracy	43	2.4
Primary school	1011	55.5
High school	528	29.0
Vocational or higher education	241	13.2
3. Environmental factors		
Breast feeding		
Less than 6 months	465	25.5
6 months or more	1358	74.5
Socioeconomic*		
Poverty	628	34.5
Borderline poverty	703	38.4
Non poverty	492	26.9

* Poverty = 5,000 baht/month, Borderline poverty = 5,001-12,000 baht/month

Non poverty = more than 12,000 baht/month

When exploring the outcome variable, the state of having suspected speech delay (SSD) was intermittent, and the sequence of the three successive states of SSD differs among children. This means that a child could be identified as SSD or normal in any wave of assessment regardless of the result of the previous or next wave assessment. The frequency distributions of the patterns of SSD for children are shown in Table 9.

There were 740 (40.6%) children who were identified as SSD cases at baseline (age 12 months). Almost half of the children (42.8%) never had any signs of SSD at any wave of the study whereas only 29 children (1.6%) had SSD at every wave of assessment. In addition, 97 (5.3%) children were identified as having suspected speech delay at age 24 months with the results of two previous assessment of speech delay indicated as normal. Overall, about half of the children (1,043 or 57.2%) were identified

as having suspected speech delay at least once. Of these, 859 (47.1%) children had normal speech development at 24 months, while 408 (22.4%) of children were classified as normal at both 18 and 24 months of age.

Table 9 Summary statistics on the result of speech delay assessment across three time points of follow-up (Complete data, N =1823)

12 months	18 months	24 months	Number (Percentage)
N	N	N	780 (42.8)
N	N	D	97 (5.3)
N	D	N	176 (9.7)
N	D	D	30 (1.6)
D	N	N	408 (22.4)
D	N	D	28 (1.5)
D	D	N	275 (15.1)
D	D	D	29 (1.6)

D = suspected speech delay case, N = normal case

The prevalence of SSD at each age based on complete data is presented in Table 10 and the trend of SSD can be seen in Figure 1. The overall prevalence in percent of children indicated as suspected speech delay is quite high (40.6%) at baseline (at 12 months). However, the prevalence with SSD substantially decreased in the following assessments. The proportion decreased from 40.6% to 28.0% for the first 6 months period of follow-up (from age 12 months to age 18 months) and dropped from 28.0% to 10.1% for the period of children aged at 18 months to 24 months. These figures demonstrate that as children grew up, fewer children were identified as having SSD. The trend of SSD prevalence is also shown in Figure 1.

Table 10 Number and percentage of SSD cases at each time point (Complete data)

Time points of assessment	Number (Percent)
At 12 months	740 (40.6%)
At 18 months	510 (28.0%)
At 24 months	184 (10.1%)

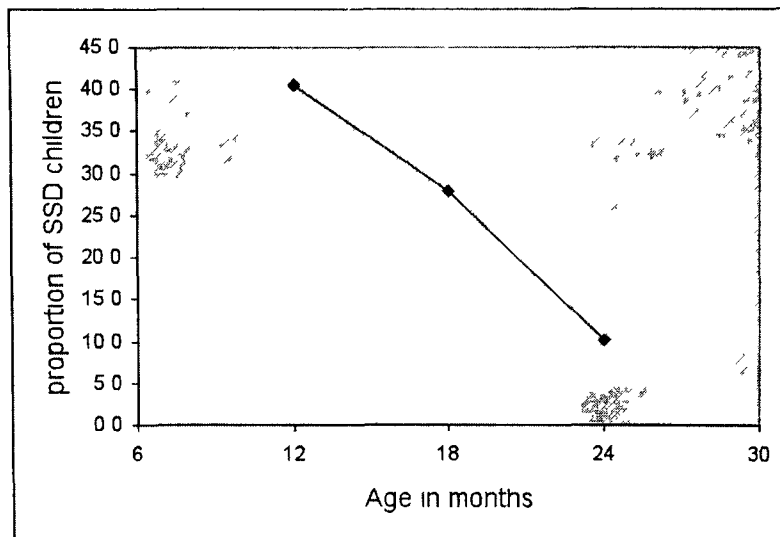


Figure 1 Proportions of children defined as having SSD at each age group

4.2 Longitudinal Data Modeling for Predicting SSD Based on Complete Data

In this part, the complete data set was used for building multivariable models for predicting SSD. To identify risk factors related to SSD in children, bivariate analyses based on ordinary logistic regression were carried out in order to examine the relationship between each independent and outcome variable. Those independent variables that were shown to be statistically significant (candidate risk factors) were entered into the multivariable model selection for obtaining the final model. The process of variable selection was based on the likelihood ratio test. The interactions of each variable with age (first order interactions) were also tested for inclusion in the final model

The models based on ordinary logistic regression, GEE and random effects models were employed to determine the risk factors associated with having SSD across all three time points. For all models, the dependent variables are the repeated assessment

measures of suspected speech delay status in children at 12, 18 and 24 months. The independent variables include age, gender, low birth weight, apgar score, birth order, maternal age, breast feeding, parental education and socioeconomics status.

As a first step of the modeling process, a series of the bivariate analyses were carried out by applying ordinary logistic regression to examine individual associations between covariates and suspected speech delay. Two classes of covariates were considered for analysis. Age of children (12, 18 and 24 months) is the time-varying covariate in this study. Time-invariant covariates included gender, low birth weight, apgar score, birth order, maternal age, breast feeding, father's education, mother's education and socioeconomic status. (see appendix A for the table of variable description and coding). The outcome variable is speech development which is categorized as "normal" and as "suspected speech delay"

According to the descriptive statistics, the impact of children's age on SSD was obvious. Thus, the analysis began by looking at the effect of age on SSD. Age was considered as a categorical variable. It was coded as 0, 1 and 2 for age at 12, 18 and 24 months, respectively. Age with this coding was further tested to determine if it could be treated as linear or if a quadratic term was needed. Next the nature of the relationship between age and SSD was investigated by examining the predicted log odds of SSD at each age category. It was found that the relationship between age and SSD was not linear (predicted log odds at 12, 18 and 24 months were -0.38, -0.95 and -2.20, respectively). The term age^2 was added in the model in order to permit nonlinear trend (quadratic).

The results of various bivariate analyses are presented in Table 11. Under preliminary investigation of all eleven candidate baseline variables related to the speech

delay, seven variables namely, age, age², gender, low birth weight, maternal age, father's and mother's education met the criteria of providing evidence of association with SSD with p-value ≤ 0.25 . These variables were considered as candidate risk factors for the multivariable models. It is noted that among these variables, age, age², father's and mother's education showed a strong relationship with p-value ≤ 0.01 . The ordinary logistic regression was used to refine the model of risk factors to predict SSD.

Table 11 Results of bivariate analysis using ordinary logistic regression based on complete data

Covariates	Coefficient (β)	SE (β)	Odds ratio	95% CI	p-value
Gender					
Female			1.00		
Male	0.278	0.062	1.32	1.17, 1.49	<0.001
Age	-0.844	0.041	0.43	0.40, 0.47	<0.001
Age²	-0.450	0.023	0.64	0.61, 0.67	<0.001
Low birth weight					
≥ 2500 gms			1.00		
< 2500 gms	0.190	0.100	1.21	0.99, 1.47	0.058
Apgar					
Normal			1.00		
Abnormal	-0.414	0.557	0.66	0.22, 1.97	0.439
Birth order					
1 st child			1.00		
2 nd child	-0.319	0.398	0.73	0.33, 1.59	0.282
3 rd child, higher	0.449	0.324	1.57	0.83, 2.95	
Maternal age	-0.007	0.005	0.99	0.98, 1.00	0.147
Paternal education					
Vocational or higher education			1.00		
High school	0.154	0.108	1.17	0.94, 1.44	<0.001
Primary school	0.375	0.099	1.45	1.20, 1.77	
Illiteracy	-0.086	0.237	0.92	0.58, 1.46	
Maternal education					
Vocational or higher education			1.00		
High school	0.065	0.162	1.07	0.78, 1.47	0.006
Primary school	-0.104	0.168	0.90	0.65, 1.25	
Illiteracy	-0.248	0.184	0.78	0.54, 1.12	

Table 11 Results of bivariate analysis using ordinary logistic regression based on complete data (Cont.)

Covariates	Coefficient (β)	SE (β)	Odd ratio	95% CI	p-value
<i>Breast feeding</i>					
< 6 months			1.00		
≥ 6 months	-0.044	0.071	0.96	0.83, 1.10	0.535
<i>Socioeconomic status</i>					
Poverty			1.00		
Borderline poverty	0.002	0.072	1.00	0.87, 1.15	0.538
Non poverty	-0.076	0.079	0.93	0.79, 1.08	

4.2.1 Fitting Ordinary Logistic Regression Model Based on Complete Data

The first multivariable model fitted to the complete data was ordinary logistic regression which assumes each repeated assessment in the same child is independent. As presented in Table 12, age which is time-varying covariate had a high impact on the probability of having SSD and the patterns of change over time had a linear and quadratic component. The older children had the lower risk of having SSD. For other time invariant covariates, gender also showed a significant effect on SSD. Male children were 1.4 times more likely to be identified as having SSD than female children. Another important predictor of having SSD was low birth weight. Children with low birth weight had a significantly greater chance of having SSD than children born with normal weight (OR = 1.3). Maternal education was not significantly related to SSD, but paternal education was. The lower level of father's education was significantly associated with higher risk of having SSD. Surprisingly, paternal literacy was a protective factor for SSD. Children with illiterate fathers were about 12% less likely to have SSD than children whose father attended vocational or higher education (OR = 0.88).

Table 12 Main effects logistic regression model predicting SSD for complete data

Covariates	Coefficient (β)	SE (β)	Odds ratio	95% CI	p-value
Constant	-0.848	0.107		-	<0.001
Age	-0.232	0.133	0.79	0.61, 1.03	0.081
Age²	-0.339	0.069	0.71	0.62, 0.82	<0.001
Gender					
Female (ref)			1.00		
Male	0.317	0.065	1.37	1.21, 1.56	<0.001
Low birth weight					
≥ 2500 gms (ref)			1.00		
< 2500 gms	0.251	0.105	1.29	1.04, 1.58	0.017
Paternal education					
Vocational or higher education (ref)			1.00		
High school	0.179	0.112	1.20	0.96, 1.49	0.110
Primary school	0.406	0.103	1.50	1.23, 1.84	<0.001
Illiteracy	-0.121	0.246	0.88	0.55, 1.44	0.623

The last step of model building was testing interaction terms. The first order interactions of each variable with age and age² were examined in the main effect model. All interaction effects were investigated for statistical significance. The significant level for testing interaction was at 0.05. As presented in Table 13, no interaction terms were significant in the model. Thus, the final logistic regression model for predicting SSD contained age, age², gender, low birth weight and father's education (presented in Table 12).

Table 13 Results of testing interaction terms to be added in main effect logistic regression model for complete data

Interaction term	Log-likelihood	Likelihood ratio test	df	p-value
Main effect only	-2881.9957			
AGE*GENDER	-2881.9585	0.07	1	0.79
AGESQ*GENDER	-2881.3798	1.23	1	0.27
AGE*LBW	-2881.6298	0.73	1	0.39
AGESQ*LBW	-2881.9933	0.005	1	0.94
AGE*FEDU	-2880.5968	2.80	3	0.43
AGESQ*FEDU	-2879.3565	5.28	3	0.15

AGE = Age, AGESQ = Age², GENDER = Gender, LBW = Low birth weight, FEDU =Father's education,

4.2.2 Fitting GEE Logistic Regression Models Based on Complete Data

A marginal or population average model was fit using two different working correlation structures: GEE model with exchangeable; GEE(exc) and autoregressive; GEE(ar1) working correlation structures. The two models were fit to the PCTC data using the same modeling strategies for building OLR model, including variables selection and testing interaction terms. Sets of interaction terms considered in GEE models are the same as those used for OLR model. They are AGE*GENDER, AGESQ*GENDER, AGE*LBW, AGESQ*LBW, AGE*FEDU and AGESQ*FEDU. The GEE technique is a semiparametric marginal modeling approach which does not specify a full probability model for the data; therefore, the traditional likelihood ratio tests for model fit and covariates significance testing are not available. Fortunately, Wald test statistics are obtainable in GEE estimation. Hence, in the variables selection process, the Wald test was used instead of the likelihood ratio test to determine significance. In fitting the two GEE models (exchangeable and autoregressive models), all interaction terms were not significant. (The results of testing interaction for the two models are shown in Appendix A: Table B1 and Table B2). The results of fitting GEE models with

exchangeable and autoregressive correlation are displayed in Table 14 and Table 15, respectively.

Table 14 GEE logistic model (exchangeable) predicting SSD for complete data

Covariates	Coefficient (β)	SE (β)	Odd ratio	95% CI	p-value
<i>Constant</i>	-0.851	0.110		-	<0.001
<i>Age</i>	-0.232	0.117	0.79	0.63, 1.00	0.046
<i>Age²</i>	-0.339	0.064	0.71	0.63, 0.81	<0.001
<i>Gender</i>					
Female (ref)			1.00		
Male	0.320	0.071	1.38	1.20, 1.58	<0.001
<i>Low birth weight</i>					
≥ 2500 gms (ref)			1.00		
< 2500 gms	0.252	0.115	1.29	1.03, 1.61	0.028
<i>Paternal education</i>					
Vocational or higher Education (ref)			1.00		
High school	0.174	0.118	1.19	0.94, 1.50	0.140
Primary school	0.408	0.107	1.50	1.22, 1.85	0.001
Illiteracy	-0.124	0.271	0.88	0.52, 1.50	0.648

Table 15 GEE logistic model (autoregressive) predicting for complete data

Covariates	Coefficient (β)	SE (β)	Odd ratio	95% CI	p-value
<i>Constant</i>	-0.846	0.110			<0.001
<i>Age</i>	-0.233	0.117	0.79	0.63, 1.00	0.045
<i>Age²</i>	-0.339	0.064	0.71	0.63, 0.81	<0.001
<i>Gender</i>					
Female (ref)			1.00		
Male	0.312	0.071	1.37	1.19, 1.57	<0.001
<i>Low birth weight</i>					
≥ 2500 gms (ref)			1.00		
< 2500 gms	0.239	0.114	1.27	1.02, 1.59	0.037
<i>Paternal education</i>					
Vocational or higher Education (ref)			1.00		
High school	0.181	0.118	1.20	0.95, 1.51	0.123
Primary school	0.408	0.107	1.50	1.22, 1.85	<0.001
Illiteracy	-0.131	0.271	0.88	0.52, 1.49	0.629

As presented in Table 14 and Table 15, both models showed a similar set of significant predictor variables. They are age, age², gender, low birth weight and father's education. When comparing the two GEE models and OLR models, an important difference was the linear age effect was significant in the two GEE models (p-value = 0.046 for exchangeable correlation and 0.045 for autoregressive correlation) but not the OLR models (p-value = 0.081; results from Table 12). However, the parameter estimates for the linear age effect were similar for OLR model and GEE(exc) model (-0.232), and were very close for OLR model and GEE(ar1) model (-0.232 and -0.233, respectively). For the quadratic age effect, the parameter estimates from both GEE models were very close [-0.340 for GEE(exc) model and -0.399 for GEE(ar1) model], and the standard errors from both GEE models are identical at three decimal place (0.064). For other covariates, gender, low birth weight and father's education, the regression coefficients were slightly different and standard errors were almost identical for both GEE models.

Both GEE models indicate that age, age², gender, low birth weight and father's education were associated with having SSD. The linear and quadratic age effects are significantly less than zero indicating a decreasing trend of having SSD as children grow. Male toddlers were more likely to be labeled as SSD than female toddlers with odds ratio = 1.38 in GEE(exc) model and 1.37 in GEE(ar1) model. Children with low birth weight (< 2500 gm) had more chance to be defined as having SSD than children with normal birth weight, odds ratio in GEE(exc) model and GEE(ar1) model were 1.29 and 1.27, respectively. Both models also indicated that children born to illiterate fathers were approximately 12% less likely to have SSD than those who born to fathers with vocational or higher education (OR = 0.88 in both models).

Overall, the direction and magnitude of parameter estimates as well as their standard errors for GEE(exc) model were quite similar to those for GEE(ar1) model. This implies that the GEE model assuming uniform correlations across time (exchangeable correlation structure) performs as well as GEE model with autoregressive working correlation which assumes that observations are only related to their own past values through first order autoregressive process.

4.2.3 Fitting Random Effects Logistic Models Based on Complete Data

Two random effects (RE) models: a random intercept model and a random intercept and a random slope for linear age covariate model were fit to the PCTC data. The modeling strategies to identify main significant covariate effects as well as interaction terms for fitting two random effects models are the same as those for fitting ordinary logistic regression. However, fitting the random effects model is more complicated and time consuming than logistic regression. The NLMIXED procedure in SAS was used to fit the models. This procedure converged and provided estimations only when the initial parameters were close to the final solution. This study used parameter estimates obtained from OLR model as the initial parameters for fitting the model. When many covariates are put in the model, the NLMIXED procedure may fail to converge; however this problem was not encountered with the set of covariates considered on this study. The interactions were also not significant in the model. (Appendix B: Table B3 and Table B4). The results of fitting the two random effects models are displayed in Table 16 and Table 17

Table 16 Random effect logistic model (random intercept) predicting SSD for complete data

Covariates	Coefficient (β)	SE (β)	Odds ratio	95% CI	p-value
<i>Constant</i>	-0.946	0.126		-	<0.001
<i>Age</i>	-0.281	0.141	0.76	0.57, 1.00	0.046
<i>Age</i> ²	-0.359	0.073	0.70	0.61, 0.81	<0.001
<i>Gender</i>					
Female (ref)			1.00		
Male	0.352	0.078	1.42	1.22, 1.66	<0.001
<i>Low birth weight</i>					
≥ 2500 gms (ref)			1.00		
< 2500 gms	0.278	0.127	1.32	1.03, 1.69	0.029
<i>Paternal education</i>					
Vocational or higher edu. (ref)			1.00		
High school	0.193	0.133	1.21	0.93, 1.57	0.146
Primary school	0.448	0.122	1.57	1.23, 1.99	0.003
Illiteracy	-0.138	0.290	0.87	0.49, 1.54	0.635
<i>Random intercept variance</i> $\hat{\tau}_0^2$	0.584	0.121	-	-	<.001

Table 17 Random effect logistic model (random intercept and slope) predicting SSD for complete data

Covariates	Coefficient (β)	SE (β)	Odds ratio	95% CI	p-value
<i>Constant</i>	-1.306	0.174	0.27	0.19, 0.38	<0.001
<i>Age</i>	0.059	0.185	1.06	0.74, 1.52	0.752
<i>Age</i> ²	-0.557	0.117	0.573	0.46, 0.72	<0.001
<i>Gender</i>					
Female (ref)			1.00		
Male	0.427	0.098	1.53	1.26, 1.86	<0.001
<i>Low birth weight</i>					
≥ 2500 gms (ref)			1.00		
< 2500 gms	0.377	0.156	1.46	1.07, 1.98	0.016
<i>Paternal education</i>					
Vocational or higher Education (ref)			1.00		
High school	0.247	0.166	1.28	0.92, 1.77	0.136
Primary school	0.523	0.153	1.69	1.25, 2.28	<0.001
Illiteracy	-0.120	0.361	0.89	0.49, 1.80	0.739
<i>Random intercept variance</i> $\hat{\tau}_0^2$	5.254	0.919	-		<0.001
<i>Random slope variance</i> $\hat{\tau}_1^2$	2.254	0.508			<0.001

As shown in Table 16 and Table 17, the two random effects models reported the same results for significant predictors of having SSD. They are age, age², gender, low birth weight and father's education. The parameter estimates as well as standard errors of time-invariant covariates for the two random effects models are slightly different. The magnitudes of regression coefficients for the linear age covariate between these two models are different. Linear age effect was significant in a random intercept model but not in a random intercept and a random slope model; however this effect was marginally significant with p-value = 0.046. Moreover, the directions of linear age effect for these two models were opposite. It was negative in a random intercept model and was positive in a random intercept and a random slope model. The estimated parameter for quadratic age covariate differs slightly between these two models. In a random intercept and a random slope model, although age linear covariate and age quadratic covariate have different direction of association with SSD, but when combining age linear and age quadratic effect together, the overall of age effect still results in negative effect on SSD. This result can be interpreted as an overall decreasing trend of having SSD across children age.

Among the time invariant covariates, both models estimated a 42% - 53% increase in the odds of having SSD in male children compared to female children. Children born with low birth weight had a greater chance of having SSD than children born with normal birth weight (OR =1.32 for a random intercept model and 1.46 for a random intercept and a random slop model). The random intercept and random slope model indicated that children who have father with primary and high school level had about 69% and 28%, respectively more likely to have SSD than those who have father

with vocational or higher level whereas children whose fathers with illiteracy had about 11% lower risk of having SSD than children whose fathers with vocational or higher education.

To test whether adding another random slope on age linear effect in the model is better than random intercept only model, the two models were compared using the log likelihood ratio test. The value of -2 log-likelihood (-2LL) is 5727.8 for a random intercept model and 5658.7 for a random intercept and a random slope model. The likelihood ratio test equals 69.1, with a p-value < 0.001 with 2 degrees of freedom, which is significant. This means that a random intercept and a random slope model is a better model in predicting SSD than a random intercept model.

4.2.4 Comparison of Models Derived from OLR, GEE and RE Models Based on Complete Data

In this section, the regression coefficients and their standard errors are compared for all five models predicting suspected speech delay in children based on three different techniques: OLR model based on marginal approach ignoring dependency; GEE with exchangeable model and GEE with autoregressive model based on population-averaged (marginal) approach taking into account of dependency; a random intercept model and a random intercept and a random slope model based on subject specific approach. The results of the comparison among the five models are shown in Table 18.

In general, all methods provided the same conclusion of significant predictor variables for SSD. Although the differences regarding covariate effects did not change substantially across the five models, statistical tests of the linear age covariate regarding p-value are different. The statistical significance of the linear age effect was found in two

GEE models and two random effects models but not in OLR model. However, the significance of the linear age effect presented in those models was marginal with p-value ranging from 0.045 – 0.047, while the OLR model indicated the linear age effect as non-significant with p-value = 0.08. All models confirmed the quadratic age effect as a significant risk factor.

When comparing the magnitude of parameter estimates among these five models, the regression coefficients from the OLR model and the two GEE models are more similar than those from the OLR model and the two RE models. In general, regression coefficients obtained from two RE models were greater than those obtained from either OLR model or two GEE models. The random effects estimates are uniformly larger in magnitude when compare to OLR and GEE estimates. The models with a random intercept and a random slope provided the biggest regression coefficients.

As previously stated, there is a relationship between the subject-specific and marginal or population-averaged estimates, specially the ratio of the estimates from the RE model (model with a random intercept) and GEE model is approximately $(\sqrt{1 + 0.346\tau_0^2})$. The random effects estimate of τ_0^2 equals 0.584, resulting in a ratio of random effects estimates to marginal estimates of 1.10 (RE/GEE ratio). Checking the ratio for each covariate in Table 18, it was found that the difference between the coefficients vary slightly from this figure (range from 1.11 to 1.21).

When focusing on standard errors, as expected, the standard errors for time-invariant covariates such as gender, low birth weight and father's education are smaller in the OLR model, whereas the standard errors for time-varying covariates such as linear and quadratic trend are generally smaller in the two GEE models. The standard errors

from the two random effects models are larger than those from the OLR model and the two GEE models.

Table 18 Regression coefficients and standard errors obtained from fitting marginal and random effects logistic regression models to predict SSD in Thai children based on complete data

Covariates	OLR		GEE (exc)		GEE (ar1)		RE (random intercept only)		RE (random intercept and slope)		RE/GEE* ratio
	β (SE)	p-value	β (SE)	p-value	β (SE)	p-value	β (SE)	p-value	β (SE)	p-value	
Constant	-0.848 (0.106)	<0.001	-0.851 (0.110)	<0.001	-0.846 (0.110)	<0.001	-0.946 (0.126)	<0.001	-1.306 (0.174)	<0.001	1.11
Age	-0.232 (0.133)	0.080	-0.232 (0.117)	0.046	-0.233 (0.117)	0.045	-0.281 (0.141)	0.046	0.059 (0.185)	0.752	1.21
Age ²	-0.340 (0.069)	<0.001	-0.340 (0.064)	<0.001	-0.339 (0.064)	<0.001	-0.359 (0.073)	<0.001	-0.557 (0.117)	<0.001	1.06
Gender	0.317 (0.065)	<0.001	0.320 (0.071)	<0.001	0.320 (0.071)	<0.001	0.352 (0.078)	<0.001	0.427 (0.098)	<0.001	1.10
Low birth wt.	0.251 (0.105)	0.017	0.252 (0.115)	0.028	0.239 (0.114)	0.037	0.278 (0.127)	0.029	0.377 (0.156)	0.016	1.10
Father's edu. High school	0.178 (0.112)	0.111	0.174 (0.118)	0.140	0.181 (0.118)	0.123	0.193 (0.133)	0.146	0.247 (0.166)	0.136	1.11
Primary school	0.405 (0.103)	<0.001	0.408 (0.107)	0.001	0.408 (0.106)	0.001	0.448 (0.122)	<0.001	0.523 (0.153)	<0.001	1.10
Illiteracy	-0.121 (0.246)	0.622	-0.124 (0.271)	0.648	-0.131 (0.271)	0.629	-0.138 (0.290)	0.635	-0.120 (0.361)	0.739	1.11
Random intercept variance $\hat{\tau}_0^2$							0.584 (0.121)	<0.001	5.254 (0.919)	<0.001	
Random slope variance $\hat{\tau}_1^2$									2.254 (0.508)	<0.001	

* The ratio of the random effects estimate (random intercept only model) to the GEE estimate (GEE with exchangeable model)

4.3 Longitudinal Data Modeling for Predicting SSD Based on Complete at Baseline Data

In the previous section, final multivariable models based on three techniques using the complete data set were obtained. Since the complete data has a very strict criterion that only children with no missing data can be included in the analysis, it causes many children to be excluded from the analysis, and consequently leads to a much smaller sample than the initial birth cohort. Analysis including more children may be more efficient and less biased. In this section, the restriction of selecting cases for the analysis was more relaxed. The criteria for cases included in the analysis is, the children who have no missing value and children who have at most one missing value on either outcome of SSD at 18 or 24 months. This data set is called “complete at baseline” The sample size for complete at baseline data set is 2,925. The strategies of model building are the same as those used in complete case data set.

The results of bivariate analysis are presented in Table 19. Age, age², gender, low birth weight, paternal and maternal education had p-value less than 0.25. These variables are the same as those obtained from bivariate analysis using complete data except for maternal age.

Table 19 Results of bivariate analysis using ordinary logistic regression for complete at baseline data

Covariates	Coefficient (β)	SE (β)	Odd ratio	95% CI	p-value
Gender					
Female			1.00		
Male	0.349	0.052	1.41	1.28, 1.57	<0.001
Age	-0.747	0.035	0.47	0.44, 0.51	<0.001
Age²	-0.410	0.019	0.66	0.64, 0.69	<0.001
Low birth weight					
≥ 2500 gms			1.00		
< 2500 gms	0.137	0.089	1.15	0.96, 1.37	0.126
Apgar					
Normal			1.00		
Abnormal	-0.123	0.419	1.13	0.50, 2.57	0.771
Birth order					
1 st child			1.00		
2 nd child	-0.300	0.340	0.74	0.38, 1.44	0.265
3 rd child, higher	0.404	0.293	1.50	0.84, 2.65	
Maternal age	-0.003	0.004	0.99	0.99, 1.00	0.444
Paternal education					
Vocational or higher education			1.00		
High school	0.021	0.080	1.02	0.87, 1.20	0.001
Primary school	0.219	0.073	1.23	1.07, 1.42	
Illiteracy	0.324	0.170	1.38	0.99, 1.93	
Maternal education					
Vocational or higher education			1.00		
High school	0.063	0.083	1.07	0.91, 1.25	0.006
Primary school	0.204	0.076	1.23	1.06, 1.42	
Illiteracy	0.462	0.136	1.59	1.22, 2.07	
Breast feeding					
< 6 months			1.00		
≥ 6 months	0.063	0.057	1.06	0.95, 1.19	0.265
Socioeconomic status					
Poverty			1.00		
Borderline poverty	-0.071	0.064	0.93	0.82, 1.06	0.502
Non poverty	-0.058	0.064	0.94	0.83, 1.07	

All indicated candidate variables obtained from bivariate analyses were further examined for inclusion in a multivariable model. The analyses of fitting model based on three different techniques were individually discussed below.

4.3.1 Fitting Ordinary Logistic Regression Model Based on Complete at Baseline

Data

Ordinary logistic regression was again fit to complete at baseline data. Interaction terms were also tested, and no interactions reached the statistical level of 0.05 for inclusion in the model (Appendix B: Table B5). Regarding results in Table 20, age (quadratic form), gender, low birth weight and father's education showed a significant association with SSD. For age (linear form), it was considered to be included in the model even though the p-value is greater than 0.05. The model in Table 20 indicated that the risk of having SSD was linearly and quadratically related to age. Boys were 1.45 times higher risk of having SSD than girls. Children with normal weight at birth were 1.24 times more likely to be classified as SSD than children born with low birth weight. Results also indicated that higher father's education was related to lower chance of having SSD. The risk of SSD in children of father with illiteracy was about 1.5 times higher than children with vocational or higher education. It is noted that all significant risk factors indicated from the analysis are similar to those obtained from complete data.

Table 20 Logistic regression model predicting SSD for complete at baseline data

Covariates	Coefficient (β)	SE (β)	Odds ratio	95% CI	p-value
Constant	-0.923	0.077			<0.001
Age	-0.062	0.114	0.94	0.75, 1.17	0.586
Age²	-0.387	0.060	0.68	0.60, 0.76	<0.001
Gender					
Female (ref)			1.00		
Male	0.373	0.054	1.45	1.31, 1.62	<0.001
Low birth weight					
≥ 2500 gms (ref)			1.00		
< 2500 gms	0.215	0.093	1.24	1.03, 1.49	0.021
Paternal education					
Vocational or higher (ref)			1.00		
High school	0.104	0.083	1.10	0.94, 1.31	0.209
Primary school	0.324	0.076	1.38	1.19, 1.60	<0.001
Illiteracy	0.390	0.176	1.48	0.05, 2.09	0.027

4.3.2 Fitting GEE Logistic Regression Models Based on Complete at Baseline Data

A marginal logistic regression model was fit using two different assumptions of correlation structure; exchangeable and autoregressive working correlation structures. The variable selection process for GEE multivariable model is the same as for ordinary logistic regression model. Bivariate analysis was performed and testing including testing the interaction terms. No interaction terms were significant in the models (Appendix B: Table B6 and Table B7). The results of final models based on GEE approach are presented in Table 21 and Table 22, respectively. In general, the two GEE models based on different working structures indicated the same conclusion of significant risk factors for SSD. They are age, age², gender, low birth weight, father's education. In addition, the regression coefficients were very close as well as their standard errors were almost identical.

Boys were about 50% (OR = 1.46 for GEE(exc) and OR = 1.45 for GEE(ar1)) more likely to identified as SSD than girls. Children with low birth weight had higher risk of SSD than children who had been born with normal birth weight with OR = 1.24 and 1.22 in GEE(exc) and GEE(ar1), respectively. Results also indicated that higher paternal education was related to lower risk of having SSD. Both GEE models confirmed that children whose fathers with illiterate education level were 50% more likely to have SSD than children whose father have vocational or higher education (OR = 1.5).

Table 21 GEE logistic model (exchangeable) predicting SSD for complete at baseline data

Covariates	Coefficient (β)	SE (β)	Odds ratio	95% CI	p-value
<i>Constant</i>	-0.926	0.079		-	<0.001
<i>Age</i>	-0.073	0.102	0.93	0.73, 1.13	0.470
<i>Age</i> ²	-0.381	0.056	0.68	0.57, 0.79	<0.001
Gender					
Female (ref)			1.00		
Male	0.378	0.059	1.46	1.34, 1.58	<0.001
Low birth weight					
≥ 2500 gms (ref)			1.00		
< 2500 gms	0.215	0.101	1.24	1.04, 1.44	0.033
Paternal education					
Vocational or higher Education (ref)			1.00		
High school	0.100	0.087	1.11	0.93, 1.28	0.251
Primary school	0.324	0.079	1.38	1.23, 1.54	<0.001
Illiteracy	0.408	0.202	1.50	1.11 1.90	0.043

Table 22 GEE logistic model (autoregressive) predicting SSD for complete at baseline data

Covariates	Coefficient (β)	SE (β)	Odds ratio	95% CI	p-value
<i>Constant</i>	-0.919	0.079			<0.001
<i>Age</i>	-0.086	0.102	0.92	0.72, 1.12	0.399
<i>Age</i> ²	-0.375	0.056	0.69	0.58, 0.80	<0.001
Gender					
Female (ref)			1.00		
Male	0.372	0.058	1.45	1.34,1.56	<0.001
Low birth weight					
≥ 2500 gms (ref)			1.00		
< 2500 gms	0.201	0.100	1.22	1.03,1.42	0.045
Paternal education					
Vocational or higher Education (ref)			1.00		
High school	0.103	0.088	1.11	0.94, 1.28	0.236
Primary school	0.321	0.080	1.38	1.22, 1.54	<0.001
Illiteracy	0.403	0.202	1.50	1.10, 1.89	0.046

4.3.3 Fitting Random Effect Logistic Regression Models Based on Complete at Baseline Data

Subject-specific model: a random intercept model and a random intercept and a random slope model were the last two multivariable models fit to the full data. The estimated parameters obtained from OLR were used as the initial values for fitting the two random effects models in SAS program. The estimations for the two models converged. No interaction terms were significant in the models (Appendix B: Table B8 and Table B9). As presented in Table 23 and Table 24, the two models provided the same significant risk factors including age, age², gender, low birth weight and father's education. Age linear effect was negative in a random intercept model, and was positive in a random intercept and slope model. For age quadratic effect, it was positive in both models. When we consider the age linear effect together with the age quadratic effect in order to assess the nature of the overall age effect, it was negatively associated with SSD which indicated a decreasing trend of SSD in children across time. In general, the regression coefficients and standard errors of risk factors in these two models are slightly different. Both models showed that boys were about 52 – 66% more likely to be labeled as SSD than girls. Children with low birth weight history were 1.26-1.34 times higher risk of having SSD than children with normal birth weight history. The models also indicated that higher paternal education was related to lower risk of having SSD. The two models were also evaluated for a better model to predict SSD. The test showed that adding a random slope to the model made a significant improvement in predicting SSD. The statistic test is significant with the likelihood ratio test = 75.3 (-2LL for random intercept model and random intercept and random slope model = -8145.9 and -8070.6, respectively).

Table 23 Random effect logistic model (random intercept) predicting SSD for incomplete data

Covariates	Coefficient (β)	SE (β)	Odds ratio	95% CI	p-value
<i>Constant</i>	-1.030	0.090	-		<0.001
<i>Age</i>	-0.110	0.121	0.90	0.66, 1.13	0.360
<i>Age</i> ²	-0.402	0.064	0.67	0.54, 0.79	<0.001
Gender					
Female (ref)			1.00		
Male	0.417	0.065	1.52	1.39, 1.64	<0.001
Low birth weight					
≥ 2500 gms (ref)			1.00		
< 2500 gms	0.233	0.111	1.26	1.04, 1.48	0.036
Paternal education					
Vocational or higher edu. (ref)			1.00		
High school	0.111	0.097	1.12	0.93, 1.31	0.253
Primary school	0.355	0.089	1.43	1.25, 1.60	<0.001
Illiteracy	0.452	0.210	1.57	1.16, 1.98	0.031
Random intercept variance $\hat{\tau}_0^2$	0.587	0.105	-		<.001

Table 24 Random effect logistic model (random intercept and slope) predicting SSD for incomplete data

Covariates	Coefficient (β)	SE (β)	Odds ratio	95% CI	p-value
<i>Constant</i>	-1.420	0.132			<0.001
<i>Age</i>	0.289	0.163	1.34	1.02, 1.65	0.047
<i>Age</i> ²	-0.620	0.108	0.54	0.33, 0.75	<0.001
Gender					
Female (ref)			1.00		
Male	0.504	0.083	1.66	1.49, 1.82	<0.001
Low birth weight					
≥ 2500 gms (ref)			1.00		
< 2500 gms	0.293	0.138	1.34	1.07, 1.61	0.029
Paternal education					
Vocational or higher Education (ref)			1.00		
High school	0.129	0.123	1.14	0.90, 1.38	0.146
Primary school	0.402	0.113	1.49	1.27, 1.72	0.003
Illiteracy	0.526	0.263	1.69	1.18, 2.21	0.632
Random intercept variance $\hat{\tau}_0^2$	4.712	0.758	-		<.001
Random slope variance $\hat{\tau}_1^2$	2.033	0.438	-		1.00

4.3.4 Comparison of Models Derived from OLR, GEE and RE Models Based on Complete Data

The results of the five models from complete at baseline data were also compared. As presented in Table 25, the findings regarding risk factors significantly associated with SSD from the five models are similar. They are age, age², gender, low birth weight and father's education. The parameter estimates from the two random effects models are larger than those from OLR and the two GEE models. In addition, the OLR and the two GEE parameter estimates were comparable to each other. The GEE with autoregressive structure provided the smallest parameter estimates when compare with the other four models. The regression coefficients and standard errors obtained from GEE(exc) and GEE(ar1) were slightly different. Regression coefficients from a random intercept model are smaller than those from a random intercept and a random slope model. Furthermore, age linear effect in random intercept model is negative (-0.110), whereas it becomes positive (0.289) in a random intercept and a random slope model. The RE/GEE ratio shown in Table 25 confirmed that the relation between random effects (random intercept model) and GEE (exchangeable model) estimates as about $(\sqrt{1 + 0.346\tau_0^2})$, which is equal to $(\sqrt{1 + 0.346(0.587)}) = 1.09$. As seen in the column, estimated ratios for the risk factors were ranged from 1.10 to 1.50.

When considering standard errors, although the OLR and the two GEE parameter estimates are very close, but their standard errors are different. The standard errors of time-varying covariate: age (linear and quadratic) for OLR model are larger than those for the two GEE models, and the standard errors of time-invariant covariates: gender, low birth weight and father's education in OLR are smaller than those for the two GEE

models. However, the two random effects models approach provided the largest in both regression coefficients and standard errors than OLR and GEE models.

Table 25 Regression coefficients and standard errors obtained from fitting marginal and random effects logistic regression models to predict SSD in Thai children based on complete at baseline data

Covariates	OLR		GEE (exc)		GEE (ar1)		RE (random intercept only)		RE (random intercept and slope)		RE/GEE ratio
	β (SE)	p-value	β (SE)	p-value	β (SE)	p-value	β (SE)	p-value	β (SE)	p-value	
Constant	-0.923 (0.077)	<0.001	-0.926 (0.079)	<0.001	-0.919 (0.079)	<0.001	-1.030 (0.090)	<0.001	-1.420 (0.132)	<0.001	1.11
Age	-0.062 (0.114)	0.586	-0.073 (0.102)	0.470	-0.086 (0.102)	0.399	-0.110 (0.121)	0.360	0.289 (0.163)	0.076	1.50
Age²	-0.387 (0.060)	<0.001	-0.381 (0.056)	<0.001	-0.375 (0.056)	<0.001	-0.402 (0.064)	<0.001	-0.620 (0.108)	<0.001	1.05
Gender	0.373 (0.054)	<0.001	0.378 (0.059)	<0.001	0.372 (0.058)	<0.001	0.417 (0.065)	<0.001	0.504 (0.083)	<0.001	1.10
Low birth wt.	0.215 (0.093)	0.021	0.215 (0.101)	0.033	0.201 (0.100)	0.045	0.233 (0.111)	0.036	0.293 (0.138)	0.034	1.08
Father's edu.											
High school	0.104 (0.083)	0.209	0.100 (0.087)	0.251	0.103 (0.088)	0.236	0.111 (0.097)	0.253	0.129 (0.123)	0.295	1.10
Primary school	0.324 (0.076)	<0.001	0.324 (0.079)	0.001	0.321 (0.080)	<0.001	0.355 (0.089)	<0.001	0.402 (0.113)	<0.001	1.10
Illiteracy	0.390 (0.176)	0.027	0.408 (0.202)	0.043	0.403 (0.202)	0.046	0.452 (0.210)	0.031	0.526 (0.263)	0.046	1.11
Random intercept variance τ_0^2							0.587 (0.105)	<0.001	4.712 (0.758)	<0.001	
Random slope variance τ_1^2							-	-	2.033 (0.438)	<0.001	

* The ratio of the random effects estimate (random intercept only model) to the GEE estimate (GEE with exchangeable model)

4.4 Comparison Results of Fitting Models from the Complete Data Set and the Complete at Baseline Data Set

In general, the results regarding parameter estimates and standard errors from the five models in both data sets are similar, that is, the models included the same set of risk factors. They are age, gender, low birth weight and father's education. The parameter estimates and their standard errors from the two RE models are uniformly greater than those OLR and GEE models. OLR and GEE parameter estimates were very close but their standard errors were different. The regression coefficients and standard errors from GEE with exchangeable and with autoregressive structure models are rather similar. One obvious difference can be seen in the results between complete data and complete at baseline data is the effect of father's education, the complete at baseline data analysis indicated that lower level of father's education was associated with higher risk of SSD; illiteracy is the highest risk of having SSD when compared to other higher education levels (primary school, high school and vocational or higher level). In contrast, the complete data analysis showed that children whose fathers did not attend school had a lower risk of SSD than children born to father who attained vocational or higher education.

CHAPTER V

DISCUSSION AND CONCLUSIONS

This final chapter begins with a discussion of the findings of identifying risk factors associated with SSD in Thai children using the PCTC data. The implications of ignoring the time dependency in longitudinal studies while modeling binary responses are described. Then the results of comparison of different statistical methods in building models to predict SSD including the relationship between marginal and subject-specific regression models are addressed. Next is a discussion of choosing an appropriate statistical method for analyzing correlated data as well as interpretation of regression coefficients followed a discussion of the influence of missing data on regression parameters. Finally, some conclusions from current the study and recommendations are addressed.

5.1 Identification of Factors Associated with Suspected Speech Delay

The general conclusion concerning risk factors related to SSD is the same for all final models based on the different approaches in both data sets (complete data and complete at baseline data). The significant risk factors are age (linear and quadratic effect), gender, low birth weight and paternal education. Among these risk factors, the effect of the time-varying child age predictor (age linear effect and age quadratic) is inverse. This finding supported the literatures which indicated that one of the most important developmental language periods is though to cover 8 month through 2 years of age (Tomasello, 2003), and approximately 15% of children at 2 years of age was

reported as late talking and resolves in 40 % to 60% of children by 3 years (Zeger, Liang, & Albert, 1991). In this study, risk of SSD declined as children grew. The prevalence of SSD at 12, 18 and 24 months are 40.6%, 28.9% and 10.1% respectively). Most children who are classified as SSD are able to achieve normal speech skills by the third wave assessment (24months), given that the prevalence of SSD drops remarkably to only 10.1% compare with the 40.6% at baseline. For time-invariant risk factors, the findings from the current study showed that male sex and low birth weight significantly increase the risk of speech delay. These results are consistent with previous studies reporting differences in the speech production and language abilities of children according to gender (Campbell, 2000; Choudhury & Benasich, 2003; Reilly et al., 2007; Shriberg, Tomblin, & McSweeny, 1999), low birth weight (Weindrich, Jennen-Steinmetz, Laucht, Esser, & Schmidt, 1998; Yliherva, Olsn, Maki-Torkko, Koironen, & Jarvelin, 2001). In this study, the effect of paternal education was slightly different between the two data sets. In complete data, the effect of father's education is not straightforward. When considering vocational or higher degree as a reference, the lower education level (primary and high school) was found to increase the risk of SSD but not illiteracy. In complete at baseline data, an inverse association between father's education and the risk of SSD was observed. One possible explanation of this difference is that more data were used in complete at baseline data analysis, and thus the estimates regression coefficients from complete at baseline data seem sensible and less biased than those from complete data. This significant effect of father's education was supported by the previous study (Tomblin, Smith, & Zhang, 1997).

Generally, the risk factors found to be related to SSD from this study are consistent with much of the published literatures except for parental education. High paternal education (vocational degree and higher) was unclear as a protective effect of speech delay. Maternal education was a significant factor consistently reported in the literature to identify children at risk, but it was not significant in this study. It would be beneficial to conduct longitudinal study on older age of children (for example, more than 24 months). It would also be interesting to evaluate speech delay over several periods of time in case and control study in stead of being study in one group.

This is the first longitudinal study undertaken in Thailand to examine concurrently several of variables affecting language development. This study adds to the growing body of knowledge to support risks for speech delay discussing in the literatures, particularly for a developing country, and it provides directions for further research concerning the biologic, familial, sociodemographic aspects that may provoke delays in speech development in Thai children.

5.2 Comparison of Marginal Models and Subject-specific Models

This section reviews the results of the comparison of the parameter estimates and their standard errors from two statistical methods, namely population-averaged or marginal and subject-specific approach. There are five models, OLR model, two GEE models assuming exchangeable and autoregressive correlation structures and two RE models; a random intercept model, and a random intercept and a random slope with age linear effect model were compared. The first three models are based on the marginal

approach, while the last two are based on the subject-specific approach. The different models were applied to complete data and complete at baseline data.

To evaluate the implication of ignoring the dependency of the observations while modeling binary responses in a longitudinal study, the results of the OLR models can be compared to the GEE and RE models which correct for the dependency of observations. Although the magnitude of the regression coefficients and their standard errors slightly differed between OLR and GEE models, the major differences are observed between OLR and RE models. According to the study of Zeger, when the sample size is large and missing data are not an issue, such as no missing data or data missing completely at random (MCAR), the estimated parameters from OLR model should be very similar to the estimated from GEE method. However, the standard errors from the OLR model are biased (Zeger, 1988). In general, failure to adjust for the dependency of the repeated observations leads to an underestimation of the standard errors of the time-invariant covariates and to an overestimation of the standard errors of the time-varying covariates (Zeger & Liang, 1986). Consistent with the statistical literature (Dunlop, 1994; Fitzmaurice, Laird, & Rotnitzky, 1993; Kuchibhatla & Fillenbaum, 2003; Zeger & Liang, 1986), the results of analyses for both complete data and complete at baseline data showed that for time-varying covariate age, the standard errors from OLR models were higher than those from the two GEE models. For time-invariant covariates gender, low birth weight and paternal education, the standard errors from OLR models were smaller than those from the two GEE models.

Both GEE and RE models techniques are considered suitable for analyses of longitudinal data, because in both techniques, a correction is made for the dependency of

the repeated observations within the same subject. With the GEE method, the correction for the dependency of observations is made by assuming a certain working correlation structure for the repeated measurements of outcome, while the RE model method allows the relationship between the outcome and the covariates to differ between subjects due to the addition of the random effect (μ_i) which are assumed to vary independently from one subject to another according to a common distribution (normal distribution). The regression coefficients calculated from the GEE method represent the average value of the individual lines or called 'population-averaged'. The regression coefficients calculated from RE models are called 'subject-specific' because the coefficients (slopes and intercepts) are allowed to be random. In case of continuous outcome variables, both GEE and RE method (GEE with exchangeable correlation structure and a random intercept model) provide exactly the same estimated regression coefficients and their standard errors. For binary outcome variables; however, the analysis is more complicated than continuous variables. In logistic regression analysis, both methods provide different results. This is the fact that in a linear model the marginal effect, or average difference for subpopulations classified by different covariate values, are the same as expected differences for individual subjects with different covariate values. However, this concept cannot be applied to the logistic model, for which a nonlinear link function is used to provide realistic connection between a linear covariate and the mean of the probability of the outcome (Diggle, Liang, & Zeger, 2002; Neuhaus, Kalbfleisch, & Hauch, 1992).

The parameter estimates from the random effects models (a random intercept model) are generally larger than those from GEE models (Neuhaus, Kalbfleisch, &

Hauch, 1991). The differences in the estimates between the two approaches are based on the correlation between the repeated observations or inter-individual heterogeneity which can be assessed through the intercept and slope variance in the RE models (Carriere & Bouyer, 2002). The results of analyses in both data sets are consistent with previous studies (Ananth, Platt, & Savitz, 2005; Crouchley & Davies, 1999; Hu, Goldberg, Hedeker, Flay, & Pentz, 1998; Kuchibhatla & Fillenbaum, 2003) that the regression coefficients and their standard error of covariates from a random intercept model are larger than those from the two GEE models (the results can be seen in Table 18 and Table 25).

5.3 Relationship between Marginal and Subject-specific Models

The estimates from the random effects models are generally larger than those from GEE method (Neuhaus, Kalbfleisch, & Hauch, 1991). According to Zeger and Liang, a regression coefficient from marginal logistic model (β_{PA}) can be approximated from the regression coefficient from the subject-specific logistic model (β_{SS}) (Zeger, Liang, & Albert, 1988) as:

$$\beta_{PA} \approx \frac{\beta_{SS}}{\sqrt{1 + 0.346(\tau_0^2)}}$$

Where τ_0^2 is random intercept variance

This relationship can be addressed by checking the estimates obtained from these two methods; the GEE with exchangeable model and a random intercept model listed in Table 18 and Table 25. For example, consider the regression coefficients of gender in the two models in Table 18. The $\hat{\beta}_{SS(\text{gender})}$ is 0.352. The estimated variance of random

intercept ($\hat{\tau}_0^2$) is 0.584. The $\hat{\beta}_{PA(gender)}$ is 0.320, which is approximately equal $0.352 / \sqrt{1 + 0.346(0.584)}$. Other covariates can also be verified in the same manner.

5.4 Interpretation of the Regression Coefficients

The interpretation of the estimated parameters obtained from either GEE and RE models is not straightforward. The regression coefficients from these two statistical approaches in the context of binary responses have different interpretation. In the GEE model, the exponential of a regression parameter is a population-averaged odds ratio of SSD for children with and without the risk factor. In other words, regression coefficients derived from marginal models (GEE models) are interpreted the same way as those derived from fitting ordinary logistic regression to a cross-sectional study. In the RE model, the exponential of a regression parameter is an odds ratio of SSD for a child that has a risk factor relative to this same child if he/she were free of this risk factor.

5.5 Impact of Missing Data on Regression Inference

Missing data are a common problem concerned in data analysis. Historically, the analysis was restricted to cases with no missing data: a complete case analysis. This approach could lead to severely biased estimates of regression parameters. Little and Rubin (Little & Rubin, 2002) classified missing data according to the mechanism that generates the missing values. Three type of missing data mechanisms were described: missing completely at random (MCAR), missing at random (MAR), and not missing at random (NMAR). Data are MCAR if the missingness is completely independent of the outcome, whether missing outcomes or observed outcomes. In this case, missing subjects

can be considered as a simple random sample of all subjects. When data are MCAR, parameter estimates are unbiased; however, statistical power is still lower than if no data were missing. Missing data are MAR if the missingness is independent of missing outcomes but may depend on the observed outcomes. Finally, NMAR, or nonignorable missing data are those for which the missingness depends on other missing values. In order for the models to provide reliable parameter estimates in the presence of missing data, certain assumptions have to be satisfied. The marginal model, using the GEE technique, requires that the data are MCAR, a most strict assumption (Diggle, Liang, & Zeger, 2002). However, subject-specific models, such as random effects models, require the less restrictive assumption that data are MAR (Diggle, Liang, & Zeger, 2002).

Although the sample size of complete at baseline data ($N = 2,925$) is much larger than complete data ($N = 1,823$), the estimates of OLR, GEE and RE models from both data sets were similar. The risk factors identified were the same, except for father's education. The regression coefficients of this risk factor are different. The effect of father with illiteracy on SSD is negative (protective effect) for the complete data but is positive (risk effect) for complete at baseline data. According to the literature review of language development, the result from complete at baseline data is more plausible than that from the complete case data. The result of analysis based on complete at baseline data seems to identify risk factors with less bias than the complete case analysis does. The impact of missing data showed in this study supported a previous study which indicated that the influence of missing data in the analysis of a binary outcome variable was rather unpredictable (Twisk, 2004).

5.6 Conclusions

The nature of longitudinal design does require the consideration of the correlation among responses on the same subject at the different points of time. Failure to take this correlation into account during statistical analysis is likely to yield biased parameter estimates, thereby leading to incorrect inferences. Regression models, (OLR) ignoring time dependency tends to overestimate the standard errors of time-invariant covariates and overestimate the standard errors of the time-varying covariates. This in turn can affect the conclusions drawn from a study. Although this study found that the error of parameter estimates produced from OLR is small when compared to the GEE and RE model analyses, it is strongly recommended that methods that account for the time dependency be used. The GEE and RE models are two common approaches among available statistical methods to handle correlated data. In logistic regression analysis, these two techniques produce different parameter estimates in term of their magnitude. In general, the magnitude of the regression coefficients and standard errors calculated with RE model are always bigger than the regression coefficients calculated with GEE models. There is no clear answer for which of the two methods is better. Researchers working on longitudinal studies with binary repeated events should thoroughly understand the strengths and weaknesses of each of the approaches as well as the interpretation of regression coefficients. If one is interested in making inferences in the population average, GEE method will probably provide valid results. In contrast, if one is interested in making inferences for an individual, RE model will probably provide valid answer. However, it should also be noted that even though choosing working correlation structure is not important issue in GEE analysis in term of parameter estimates, but this

would be in case if the number of observations is large. Accurately specifying a correct working correlation structure is still be needed and this will definitely improve the efficiency of GEE estimates. Also, it should be realized that so far, the RE models approach (a model with intercept only and a model with a random intercept and a random slope with time) is limited by the assumption that the random effects have to be normally distributed.

Missing data are an important issue in analyzing longitudinal data. This study indicated that complete case analysis is not always a good choice for data analysis even though the number of subjects remaining in analysis is large. It would be recommend that when subjects have missing observations, researcher should not ignore those subjects but should attempt to address missing data and handle them properly in order to reduce bias. In addition, the GEE and RE models require different assumptions of missing data. Hence, the type of missingness should be seriously considered when choosing a statistical analysis.

APPENDIX A

Table A1 Description and coding for the variables in the analyses

<i>Variables</i>	<i>Description</i>	<i>Coding</i>
Gender	Gender	0 = Female 1 = Male
AGE	Age (month)	0 = 12 months 1 = 18 months 2 = 24 months
LBW	Low birth weight	0 = \geq 2500 gms 1 = $<$ 2500 gms
APGAR	Apgar score	0 = score \geq 7 1 = score $<$ 7
ORDER	Birth order	0 = 1 st order 1 = 2 nd order 2 = 3 rd or higher order
MAGE	Maternal age	Record as is
INCOME		0 = Non poor (5,000 baht/month) 1 = Borderline poor (5,001-12,000 baht/month) 2 = Poor ($>$ 12,000 baht/month)
BF	Breast feeding	0 = \geq 6 months 1 = $<$ 6 months
FEDU	Father's education	0 = Vocational, college or higher 1 = High school 2 = Primary school 3 = Illiteracy
MEDU	Mother's education	0 = Vocational, college or higher 1 = High school 2 = Primary school 3 = Illiteracy

APPENDIX B

Table B1 Results of testing interaction terms to be added in main effect GEE model (exchangeable) for complete data

Interaction term	Coefficient	SE(β)	Wald test	p-value
Main effect only				
AGE*GENDER	-0.023	0.085	-0.27	0.79
AGESQ*GENDER	-0.051	0.045	-1.14	0.26
AGE*LBW	0.120	0.134	0.89	0.37
AGESQ*LBW	0.005	0.070	0.07	0.94
AGE*FEDU				
AGE*FEDU1	0.050	0.148	0.34	0.74
AGE*FEDU2	-0.103	0.136	-0.75	0.45
AGE*FEDU3	0.125	0.320	0.39	0.70
AGESQ*FEDU				
AGESQ*FEDU1	0.061	0.078	0.79	0.43
AGESQ*FEDU2	-0.056	0.073	-0.77	0.44
AGESQ*FEDU3	0.030	0.171	0.17	0.86

AGE = Age, AGESQ = Age², GENDER = Gender, LBW = Low birth weight, FEDU =Father's education,

Table B2 Results of testing interaction terms to be added in main effect GEE model (Autoregressive) for complete data

Interaction term	Coefficient	SE(β)	Wald test	p-value
Main effect only				
AGE*GENDER	-0.019	0.088	-0.21	0.83
AGESQ*GENDER	-0.055	0.045	-1.21	0.23
AGE*LBW	0.124	0.137	0.91	0.36
AGESQ*LBW	-0.001	0.072	-0.02	0.98
AGE*FEDU				
AGE*FEDU1	0.045	0.151	0.30	0.76
AGE*FEDU2	-0.101	0.140	-0.73	0.47
AGE*FEDU3	0.129	0.328	0.39	0.69
AGESQ*FEDU				
AGESQ*FEDU1	0.066	0.079	0.83	0.40
AGESQ*FEDU2	-0.055	0.074	-0.75	0.46
AGESQ*FEDU3	0.026	0.176	0.15	0.88

AGE = Age, AGESQ = Age², GENDER = Gender, LBW = Low birth weight, FEDU =Father's education,

It is noted that for GEE model, Wald test was used to test for each interaction term. This is because logistic regression model based on GEE method does not provide log-likelihood.

Table B3 Results of testing interaction terms to be added in main effect random effect model with only intercept model for complete data

Interaction term	Log-likelihood	Likelihood ratio test	df	p-value
Main effect only	-2863.9015			
AGE*GENDER	-2863.8302	0.14	1	0.71
AGESQ*GENDER	-2863.1178	1.57	1	0.21
AGE*LBW	-2863.5359	0.73	1	0.39
AGESQ*LBW	-2863.9011	0.0008	1	0.98
AGE*FEDU	-2862.2566	3.29	3	0.35
AGESQ*FEDU	-2860.9146	5.97	3	0.11

AGE = Age, AGESQ = Age², GENDER = Gender, LBW = Low birth weight, FEDU =Father's education,

Table B4 Results of testing interaction terms to be added in main effect random effect model with random intercept and random slope model for complete data

Interaction term	Log-likelihood	Likelihood ratio test	df	p-value
Main effect only	-2863.9017			
AGE*GENDER	-2863.8304	0.14	1	0.71
AGESQ*GENDER	-2863.1180	1.57	1	0.21
AGE*LBW	-2863.5361	0.73	1	0.39
AGESQ*LBW	-2863.9013	0.0008	1	0.98
AGE*FEDU	-2862.2567	3.29	3	0.35
AGESQ*FEDU	-2860.9148	5.97	3	0.11

AGE = Age, AGESQ = Age², GENDER = Gender, LBW = Low birth weight, FEDU =Father's education,

Table B5 Results of testing interaction terms to be added in main effect logistic regression model for complete at baseline data

Interaction term	Log-likelihood	Likelihood ratio test	df	p-value
Main effect only	-4096.8494			
AGE*GENDER	-4096.3484	1.00	1	0.32
AGESQ*GENDER	-4095.1185	3.64	1	0.63
AGE*LBW	-4096.8331	0.03	1	0.86
AGESQ*LBW	-4096.7184	0.26	1	0.61
AGE*FEDU	-4094.6249	4.45	3	0.22
AGESQ*FEDU	-4092.8554	7.99	3	0.50

AGE = Age, AGESQ = Age², GENDER = Gender, LBW = Low birth weight, FEDU =Father's education,

Table B6 Results of testing interaction terms to be added in main effect GEE model (exchangeable) for complete at baseline data

Interaction term	Coefficient	SE(β)	Wald test	p-value
Main effect only				
AGE*GENDER	-0.071	0.071	-1.00	0.32
AGESQ*GENDER	-0.072	0.039	-1.89	0.06
AGE*LBW	0.020	0.119	0.17	0.87
AGESQ*LBW	-0.035	0.064	-0.54	0.59
AGE*FEDU				
AGE*FEDU1	-0.047	0.112	-0.42	0.68
AGE*FEDU2	-0.187	0.103	-1.82	0.07
AGE*FEDU3	-0.171	0.237	-0.72	0.47
AGESQ*FEDU				
AGESQ*FEDU1	-0.004	0.060	-0.07	0.94
AGESQ*FEDU2	-0.112	0.056	-2.01	0.05
AGESQ*FEDU3	-0.155	0.134	-1.15	0.25

AGE = Age, AGESQ = Age², GENDER = Gender, LBW = Low birth weight, FEDU =Father's education,

Table B7 Results of testing interaction terms to be added in main effect GEE model (Autoregressive) for complete at baseline data

Interaction term	Coefficient	SE(β)	Wald test	p-value
Main effect only				
AGE*GENDER	0.005	0.082	0.06	0.95
AGESQ*GENDER	-0.049	0.044	-1.11	0.27
AGE*LBW	0.056	0.133	0.42	0.68
AGESQ*LBW	-0.027	0.073	-0.37	0.71
AGE*FEDU				
AGE*FEDU1	0.034	0.130	0.27	0.79
AGE*FEDU2	-0.095	0.119	-0.80	0.43
AGE*FEDU3	0.050	0.262	0.19	0.85
AGESQ*FEDU				
AGESQ*FEDU1	0.068	0.073	0.91	0.36
AGESQ*FEDU2	-0.050	0.069	-0.73	0.47
AGESQ*FEDU3	-0.050	0.152	-0.33	0.74

AGE = Age, AGESQ = Age², GENDER = Gender, LBW = Low birth weight, FEDU =Father's education,

Table B8 Results of testing interaction terms to be added in main effect random effect model with only intercept model for complete at baseline data

Interaction term	Log-likelihood	Likelihood ratio test	df	p-value
Main effect only	-4072.3894			
AGE*GENDER	-4071.7344	1.31	1	0.25
AGESQ*GENDER	-4070.4794	3.82	1	0.05
AGE*LBW	-4072.3827	0.01	1	0.91
AGESQ*LBW	-4072.2090	0.36	1	0.55
AGE*FEDU	-4069.6505	5.48	3	0.14
AGESQ*FEDU	-4068.5194	7.74	3	0.05

AGE = Age, AGESQ = Age², GENDER = Gender, LBW = Low birth weight, FEDU =Father's education,

Table B9 Results of testing interaction terms to be added in main effect random effect model with random intercept and random slope model for complete at baseline data

Interaction term	Log-likelihood	Likelihood ratio test	df	p-value
Main effect only	-4072.3898			
AGE*GENDER	-4071.7347	1.31	1	0.25
AGESQ*GENDER	-4070.4798	3.82	1	0.05
AGE*LBW	-4072.3830	0.01	1	0.91
AGESQ*LBW	-4072.2098	0.36	1	0.55
AGE*FEDU	-4069.6498	5.48	3	0.14
AGESQ*FEDU	-4068.5198	7.74	3	0.05

AGE = Age, AGESQ = Age², GENDER = Gender, LBW = Low birth weight, FEDU =Father's education,

References

- Albert, P. S. (1999). Longitudinal data analysis (repeated measures) in clinical trials. *Statistics in Medicine*, *18*, 1707-1732.
- Ananth, C. V., Platt, R. W., & Savitz, D. A. (2005). Regression models for clustered binary responses: implications of ignoring the Intraclass correlation in an analysis of perinatal mortality in twin gestations. *Annals of Epidemiology*, *15*(4), 293-301.
- Andrew, H., Goldberg, D., Wellen, N., Pittman, B., & Struening, E. (1995). Prediction of special education placement from birth certificate data. *Journal of Preventive Medicine*, *11*, 55-61.
- Ansel, B. M., Landa, R. M., & Stark-Selz, R. E. (1994). Development and disorders of speech and language. In F. A. Oski & C. D. DeAngelis (Eds.), *Principles and practice of pediatrics* (pp. 686-700). Philadelphia: Lippincott.
- Austin, P. C. (2007). A comparison of the statistical power of different methods for the analysis of cluster randomization trials with binary outcomes. *Statistics in Medicine*, *26*, 3550-3565.
- Ballinger, G. A. (2004). Using generalized estimating equations for longitudinal data analysis. *Organizational Research Methods*, *7*, 127-150.
- Bellamy, S. L., Gibberd, R., Hancock, L., Howley, P., Kennedy, B., Klar, N., et al. (2000). Analysis of dichotomous outcome data for community intervention studies. *Statistical Methods in Medical Research*, *9*, 135-159.
- Bishop, J., Venables, W. N., & Wang, Y.-G. (2004). Analysing commercial catch and effort data from a Penaeid trawl fishery: A comparison of linear models, mixed models, and generalized estimating equations approaches. *Fisheries Research*, *70*(2-3), 179-193.
- Blum, N. J., & Baron, M. A. (1997). Speech and language disorders. In N. W. Schwartz (Ed.), *Pediatrics primary care: problem oriented approach* (pp. 845-849). St. Louis, MO: Mosby.
- Burden, V., Scott, V. M., Forge, J., & Goodyer, I. (1996). The Cambridge Language and Speech Project (CLASP): Detection of language difficulties at age 36 to 39 months. *Developmental Medicine & Child Neurology*, *38*, 613-631.
- Burton, P., Gurrin, L., & Sly, P. (1998). Extending the simple linear regression model to account for correlated response: an introduction to generalized estimating equations and multilevel mixed modeling. *Statistics in Medicine*, *17*(1261-1291).

- Campbell, M. J. (2000). Cluster randomized trials in general family practice. *Statistical Methods in Medical Research*, 9, 81-94.
- Cantwell, D. P., & Baker, L. (1991). *Psychiatric and developmental disorders in children with communication disorder* Washington, DC: American Psychiatric Press.
- Carlin, J. B., Wolf, R., Brown, C. H., & Gelman, A. (2001). A case study on the choice, interpretation and checking of multilevel models for longitudinal binary outcomes. *Biostatistics*, 2(4), 397- 416.
- Carriere, I., & Bouyer, J. (2002). Choosing marginal or random-effects models for longitudinal binary responses: application to self-reported disability among older persons [Electronic Version]. *BMC Medical Research Methodology*, 2. Retrieved November 12, 2007, from <http://www.biomedcentral.com/1471-2288/2/15>.
- Choudhury, N., & Benasich, A. A. (2003). A family aggregation study: The influence of family history and other risk factors on language development. *Journal of Speech, Language & Hearing Research*, 46(2), 261.
- Crouchley, R., & Davies, R. B. (1999). A comparison of population average and random-effect models for the analysis of longitudinal count data with base-line information. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 162(3), 331-347.
- Cusson, R. M. (2003). Factors influencing language development in preterm infants. *Journal of Obstetric, Gynecologic, & Neonatal Nursing*, 32(3), 402-409.
- Dee, D. L., Ruowei, L., Li-Ching, L., & Grummer-Strawn, L. M. (2007). Associations Between Breastfeeding Practices and Young Children's Language and Motor Skill Development. *Pediatrics*, 119, S92-S98.
- Diggle, P. J., Liang, K.-Y., & Zeger, S. L. (2002). *Analysis of longitudinal data* (2nd ed.). New York: Oxford University Press.
- Dollaghan, C. A., Campbell, T. F., Paradise, J. L., Feldman, H. M., Janosky, J. E., Pitcairn, D. N., et al. (1999). Maternal education and measures of early speech and language. *Journal of Speech, Language & Hearing Research*, 42(6), 1432-1443.
- Dunlop, D. D. (1994). Regression for longitudinal data: a bridge from least square regression. *American Journal of Statistics*, 48, 299-303.
- Eapen, V., Zoubeidi, T., & Yunis, F. (2004). Screening for language delay in the United Arab Emirates. *Child: Care, Health and Development*, 30(5), 541-549.

- Feldman, M. H. (2005). Evaluation and management of language and speech disorders in preschool children. *Pediatrics in Review*, 26, 131-141.
- Fitzmaurice, G. M. (1995). A caveat concerning independence estimating equations with multivariate binary data. *Biometrics*, 51, 309-317.
- Fitzmaurice, G. M., Laird, N. M., & Rotnitzky, A. G. (1993). Regression model for discrete longitudinal response. *Statistical Sciences*, 8, 284-309.
- Fox, A. V., Dodd, B., & Howard, D. (2002). Risk factors for speech disorders in children. *International Journal of Language & Communication Disorders*, 37(2), 117-131.
- Hedeker, D., & Gibbons, R. D. (2006). *Longitudinal data analysis*. New Jersey: John Wiley & Son, Inc.
- Heo, M., & Leon, A. C. (2005). Comparison of statistical methods for analysis of clustered binary observations. *Statistics in Medicine*, 24, 911-923.
- Horwitz, S. M., Irwin, J. R., Briggs-Gowan, M. J., Bosson Heenan, J. M., Mendoza, J., & Carter, A. S. (2003). Language delay in a community cohort of young children. *Journal of the American Academy of Child and Adolescent Psychiatry*, 42(8), 932-940.
- Hosmer, D. W., & Lemeshow, S. (2000). *Applied logistic regression*. New York: John Wiley & Son, Inc.
- Hu, F. B., Goldberg, J., Hedeker, D., Flay, B. R., & Pentz, M. A. (1998). Comparison of population-averaged and subject-specific approaches for analyzing repeated binary outcomes. *American Journal of Epidemiology*, 147(7), 694-694.
- Kuchibhatla, M., & Fillenbaum, G. (2003). Comparison of methods for analyzing longitudinal binary outcomes: cognitive status as an example. *Aging & Mental Health*, 7(6), 462-462.
- Law, J., Boyle, J., Harris, F., Harkness, A., & Nye, C. (1998). Screening for primary speech and language delay: A systematic review of the literature... Communicating the evidence: the case for speech and language therapy. Proceedings of the College's 1998 Conference, Liverpool 15-17 October 1998. *International Journal of Language & Communication Disorders*, 33, 21-23.
- Leung, A. K. C., & Kao, C. P. (1999). Evaluation and management of the child with speech delay. *American Family Physician*, 59(11), 3121.

- Lewis, B. A., Freebairn, L. A., & Taylor, H. G. (2000). Academic outcomes in children with histories of speech sound disorders. *Journal of Communication Disorders*, 33(1), 11-30.
- Liang, K. Y., & Zeger, S. L. (1986). Longitudinal data analysis using generalized linear model. *Biometrika*, 73, 13-22.
- Liang, K. Y., & Zeger, S. L. (1993). Regression analysis for correlated data. *Annual Review of Public Health*, 14, 43-68.
- Liao, J. G., & Lipsitz, S. R. (2002). A type of restricted maximum likelihood estimators of variance components in generalized linear mixed models. *Biometrika*, 89(2), 401-409.
- Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data* (Second ed.). New Jersey: John Wiley & Sons.
- Morisset, C. E., Barnard, K. E., & Booth, C. L. (1995). Toddler's language development: Sex differences within social risk. *Developmental Psychology*, 31(5), 851-865.
- Neuhauser, J. M. (2001). Assessing change with longitudinal and clustered binary data. *Annual Review of Public Health*, 22(1), 115-128.
- Neuhauser, J. M., Kalbfleisch, J. D., & Hauch, W. W. (1991). A comparison of Cluster-Specific and Population-Averaged approaches for analyzing correlated binary data. *International Statistical Review*, 59, 25-35.
- Neuhauser, J. M., Kalbfleisch, J. D., & Hauch, W. W. (1992). Statistical methods for longitudinal and clustered designs with binary responses. *Statistical Methods in Medical Research*, 1, 249-273.
- Pan, W. (2001). Akaike's information criterion in generalized estimating equations. *Biometrics*, 57, 120-125.
- Pendergast, J. F., Gange, S. J., Newton, M. A., Lindstrom, M. J., Palta, M., & Fiske, M. R. (1996). A survey of methods for analyzing clustered binary response data. *International Statistical Review*, 64, 89-118.
- Peter, T. J., Richard, S. H., Bankhead, C. R., Ades, A. E., & Strene, J. A. C. (2003). Comparison of methods for analysing cluster randomized trials: an example involving a factorial design. *International Epidemiological Association*, 32, 840-846.
- Prentice, R. L. (1986). Binary regression using an extended beta-binomial distribution, with discussion of correlation induced by covariate measurement errors. *Journal of The American Statistical Association* 81, 321-327.

- Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods*. California: Sage Publisher, Inc.
- Redmond, S. M., & Rice, M. L. (1998). The socioemotional behaviors of children with SLI: Social adaptation or social deviance? *Journal of Speech, Language & Hearing Research, 41*, 688-700.
- Reilly, S., Wake, M., Bavin, E. L., Prior, M., Williams, J., Bretherton, L., et al. (2007). Predicting language at 2 years of age: a prospective community study. *Pediatrics, 120*, e1441-e1449.
- Rescorla, L., Hadicke-Wiley, M., & Escarce, E. (1993). Epidemiological investigation of expressive language delay at age two *First Language, 13*, 5-22.
- Schwartz, E. R. (1990). Speech and language disorders. In M. Schwartz (Ed.), *Pediatrics primary care: a problem oriented approach* (pp. 696-700). St.Louis: Mosby.
- Shriberg, L. D., Tomblin, J. B., & McSweeney, J. L. (1999). Prevalence of speech delay in 6-year-old children and comorbidity with language impairment *Journal of Speech, Language & Hearing Research, 42*, 1461-1481.
- Singer, L. T., Siegel, A. C., Lewis, B., Hawkins, S., Yamashita, T., & Baley, J. (2001). Preschool language outcomes of children with history of bronchopulmonary dysplasia and very low birth weight. *Journal of Developmental & Behavioral Pediatrics, 22*, 19-26.
- Stanton-Chapman, T. L., Chapman, D. A., Bainbridge, N. L., & Scott, K. G. (2002). Identification of early risk factors for language impairment. *Research in Developmental Disabilities, 23*(6), 390-405.
- Stiratelli, R., Laird, N., & Ware, J. H. (1984). Random-effects models for serial observations with binary response. *Biometrics, 40*(4), 961-971.
- Tomasello, M. (2003). *Constructing a Language: A Usage-Based Theory of Language Acquisition*. Cambridge, MA Harvard University Press.
- Tomblin, J. B. (1990). The effect of birth order on the occurrence of developmental impairment. *British Journal of Disorders of Communication, 25*, 77-84.
- Tomblin, J. B., Hardy, J. C., & Hein, H. A. (1991). Predicting poor-communication status in preschool children using risk factors present at birth. *Journal of Speech & Hearing Research, 34*, 1096-1105.
- Tomblin, J. B., Smith, E., & Zhang, X. (1997). Epidemiology of specific language impairment: Prenatal and perinatal risk factors. *Journal of Communication Disorders, 30*(4), 325-344.

- Twisk, J. W. R. (2004). Longitudinal data analysis. A comparison between generalized estimating equations and random coefficient analysis. *European Journal of Epidemiology*, *19*, 769 - 776
- Ukoumunne, O. C., Carlin, J. B., & Gulliford, M. C. (2007). A simulation study of odds ratio estimation for binary outcomes from cluster randomized trials. *Statistics in Medicine*, *26*, 3415-3428.
- Weindrich, D., Jennen-Steinmetz, C., Laucht, M., Esser, G., & Schmidt, M. H. (1998). At risk for language disorders? Correlates and course of language disorders in preschool children born at risk. *Acta Paediatrica*, *87*(12), 1288-1294.
- Yliherva, A., Olsn, P., Maki-Torkko, E., Koiranen, M., & Jarvelin, M. R. (2001). Linguistic and motor abilities of low-birthweight children as assessed by parents and teachers at 8 years of age. *Acta Paediatrica*, *90*(12), 1440-1449.
- Zeger, S. L. (1988). Commentary *Statistics in Medicine*, *7*, 161-168.
- Zeger, S. L., Liang, K.-Y., & Albert, P. S. (1988). Models for longitudinal data: a generalized estimating equation approach. *Biometrics*, *44*(4), 1049-1060.
- Zeger, S. L., & Liang, K. Y. (1986). Longitudinal data analysis for discrete and continuous outcomes. *Biometrics*, *42*, 121-130.
- Zeger, S. L., & Liang, K. Y. (1992). An overview of methods for the analysis of longitudinal data. *Statistics in Medicine*, *11*, 1825-1839.
- Zeger, S. L., Liang, K. Y., & Albert, P. S. (1991). The interpretation of a regression coefficient. *Biometrics*, *47*, 1596-1597.
- Zorn, C. J. W. (2001). Generalized estimating equation models for correlated data: a review with application. *American Journal of Political Sciences*, *50*, 470-490.

BIOGRAPHY

Name: Natkamol Chansatitporn

Date of Birth: Aug 15, 1962

Gender: Female

Office address: Faculty of Public Health, Mahidol University

420/1 Ratchawithi Rd., Ratchathewi District, Bangkok 10400, Thailand

Current position: Assistant Professor

Education:

2004 - 2009 Tulane University School of Public Health and Tropical Medicine
New Orleans, LA., USA.

Sc.D. (Biostatistics)

1999 - 2001 University of the Philippines Manila
Manila, Philippines

M.Sc. (Epidemiology)

1990 - 1994 Mahidol University
Bangkok, Thailand

M.Sc. (Biostatistics)

1981 - 1985 Mahidol University
Bangkok, Thailand

B.Sc. (Nursing and Midwife)