The Effect of Birth Order on Infant Injury

by

Heather Van Duker

A selected project submitted to the faculty of

Brigham Young University

in partial fulfillment of the requirements for the degree of

Master of Science

Department of Statistics

Brigham Young University

April 2007

BRIGHAM YOUNG UNIVERSITY

GRADUATE COMMITTEE APPROVAL

of a selected project submitted by

Heather Van Duker

This selected project has been read by each member of the following graduate committee and by majority vote has been found to be satisfactory.

_____	_____
Date	Gilbert W. Fellingham, Chair
_____	_____
Date	Heather T. Keenan
_____	_____
Date	G. Bruce Schaalje
_____	_____
Date	Christopher Shane Reese

BRIGHAM YOUNG UNIVERSITY

As chair of the candidate's graduate committee, I have read the selected project of Heather Van Duker in its final form and have found that (1) its format, citations and bibliographical style are consistent and acceptable and fulfill university and department style requirements; (2) its illustrative materials including figures, tables, and charts are in place; and (3) the final manuscript is satisfactory to the graduate committee and is ready for submission to the university library.

Date	Gilbert W. Fellingham Chair, Graduate Committee

Accepted for the Department

Date	Scott D. Grimshaw Graduate Coordinator/Associate Department Chair

Accepted for the College

Date	Thomas W. Sederberg Associate Dean, College of Physical and Mathematical Sciences

ABSTRACT

The Effect of Birth Order on Infant Injury

Heather Van Duker

Department of Statistics

Master of Science

Pediatric injury is both common and expensive. Finding ways to prevent pediatric injury is a major public health concern. Many studies have investigated various aspects of pediatric injury, and some suggest that birth order may be an important risk factor for pediatric injury. This study further examined the relationship of birth order with pediatric injury, specifically studying the association of birth order with emergency department-attended infant injury while adjusting for other important family and individual covariates.

Data for analysis included Utah birth certificate, death certificate, and hospital emergency department datasets, which were probabilistically linked to obtain complete demographic and injury information for infants born in 1999—2002. Three groups of risk factors were defined for analysis: maternal demographics, maternal risk behaviors, and infant demographics. Two outcome variables were defined for this study, “injury event” and “severe injury event.” Data was analyzed using generalized estimating equations (GEE).

Birth order was associated with infant injury events and severe infant injury events. Birth order 4th or greater had the greatest effect for both injury outcomes. Additionally, several maternal characteristics were associated with infant injury events and severe infant injury events. In particular, maternal age and maternal smoking behavior were associated with increased infant injury risk.

This study identified two targeted populations that are well-suited to injury prevention efforts: infants born to mothers who smoke, and infants born to mothers who are young and have many other children.

ACKNOWLEDGMENTS

I would like to acknowledge my graduate committee chair, Dr. Gilbert Fellingham, my graduate committee, Dr. Heather Keenan, Dr. Bruce Schaalje, and Dr. Shane Reese, and the Associate Department Chair, Dr. Scott Grimshaw, for their help and support. Each has contributed a large piece to the completion of this project and given liberally of their time. I would especially like to acknowledge Dr. Fellingham for his exceptional patience, guidance, and sense of humor. I would also like to specifically thank Dr. Keenan for sharing her expertise and limited time so generously with me.

I would like to acknowledge the debt I owe to my colleagues at the University of Utah, particularly Dr. Lenora Olson and Stacey Knight. I have learnt much from working with them, and would like to thank them for their tireless support, encouragement, and interest in my success.

I would like to thank the Graduate Secretary, Megan Tate, for keeping me on track, providing me with all needed information, and taking on the tremendous task of editing.

Not least, I would like to thank my husband, William, for his patience and forbearance whilst I have spent countless hours working on this project.

CONTENTS

CHAPTER

1 INTRODUCTION 1

2 REVIEW OF LITERATURE 3

2.1 Infant Injury Literature 3

2.2 Generalized Linear Models 4

2.3 Generalized Linear Models for Binary Data 6

2.4 Generalized Estimating Equations 8

2.4.1 An overview of GEEs 9

2.4.2 Advantages of GEEs 10

2.4.3 Inference from GEE Models 11

2.4.4 GEEs and Convergence, Goodness-of-Fit, and Missing Data 12

2.5 Generalized Estimating Equations for Binary Data 13

2.5.1 Goodness-of-Fit Statistics for GEEs 13

2.5.2 Alternating Logistic Regressions Algorithm 15

2.6 Conclusion 15

3 METHODS 17

3.1 Data 17

3.2 Risk Factors 18

3.3 Outcomes 19

3.4 Analysis 20

4 RESULTS 23

4.1 Population Description 23

4.2 Mother and Infant Demographics 23
4.3 Maternal Risk Behaviors 25
4.4 GEE Analysis 26
4.4.1 All Infant Injuries 26
4.4.2 Severe Infant Injuries 29
5 DISCUSSION 31
5.1 Study Findings 31
5.2 Limitations 33
5.3 Strengths 34
5.4 Conclusions 34
5.5 Future Work 34
REFERENCES 36
APPENDIX A: ALL INJURIES, SAS CODE AND OUTPUT 41
APPENDIX B: SEVERE INJURIES, SAS CODE AND OUTPUT 43

TABLES

Table

4.1 Infant Demographics 24

4.2 Maternal Demographics 25

4.3 Maternal Risk Behaviors 26

4.4 P-Values and Adjusted Odds Ratios for All Infant Injury 26

4.5 Incidence Rate per 1 000 person-years (95% CI) Birth Order by Maternal Age 28

4.6 Adjusted Odds Ratios for Prenatal Care by Maternal Education 29

4.7 P-values and Adjusted Odds Ratios for Severe Infant Injury 30

FIGURES

Figure

4.1 Odds of Injury by Birth Order and Maternal Age.....27

1 INTRODUCTION

In 2002, injuries hospitalized or killed approximately 12 out of every 100 children under age 5 years in the United States (CDC 2002). In 1994, the estimated economic cost of injury for children under age 5 years was $75 billion (Danseco et al. 2000). Factoring in inflation, increasing costs of health care, and an increasing population, that figure is probably much greater today (Weiss et al. 1997) (McCaig and Burt 2001). Pediatric injury is costly and finding ways to prevent it has become a major public health concern.

A number of studies have investigated infant injury (Brenner et al. 1999) (Scholer et al. 1997) (Agran et al 2003) (Nathens et al. 2000). However, none of the studies have targeted the effect of birth order on emergency department (ED) attended fatal and nonfatal infant injury. Birth order may play an important role as a predictor of injury for a new infant. For example, an infant in a multiple-child household may have increased odds of injury compared to first-born infants because the possibility of lessened supervision becomes greater when the attention of the caregiver(s) is divided among many young children. Birth order also may help characterize types of injury. For example, a third-born infant may be more likely to choke on small parts of toys due to the presence of age-inappropriate toys in a household with multiple siblings.

The objective of this study was to investigate the relationship of selected individual and maternal factors with infant injuries. In particular, the association of birth order with infant injury was studied, while controlling for other individual, socio-economic, and family factors as detailed in the analysis section.

The analysis dataset is a compilation of data from the Utah birth certificate database, the Utah death certificate database, and the Utah hospital emergency

department database. The Utah datasets were particularly useful for this study because Utah has the highest birth rate in the United States (Sutton and Matthews 2004), as well as large families with many young preschool-age children in households. Data include all injury causes and intents; data are not excluded by intention or outcome (fatal or nonfatal).

The three databases were probabilistically linked using LinkSolv 7.0 (Strategic Matching, Inc.). Data were analyzed using PROC LOGISTIC and PROC GENMOD in SAS 9.1.3 (Cary, NC). The model was built using logistic regression on a randomly selected, non-correlated subset of the full data. Two models were built. One model was built using backward elimination, and the second model was built using forward elimination. The two models were built to examine any potential differences between variables selected by the two elimination methods. A generalized estimating equations (GEE) model for correlated binary data was fitted to the full dataset to account for the correlation within sibling groups and further reduced using backward elimination.

Characterizing the risk factors for injuries to first-, second-, third-, and fourth-or-more-born children in households may help injury-prevention educators target the issues specific to each child in a household and assist clinicians and parents in preventing the injuries that are common to their household type.

2 REVIEW OF LITERATURE

2.1 Infant Injury Literature

A number of investigators have studied pediatric injury. Of particular interest to this study are those studies which examine infants < 1 year of age and those studies which examine birth order (first-, second-, third-born) as a predictor of childhood injury. Brenner et al. (1999) used national birth and death certificate data to observe types of injury deaths and risk factors for fatal injury in infants < 1 year of age. They found that homicide, suffocation, motor vehicle crashes, and choking are the leading causes of death for this age group. These investigators also found that birth order is an important predictor of fatal injuries due to drowning, fire, and mechanical suffocation. Similarly, Scholer et al. (1997) examined linked national birth and infant death data to identify socio-demographic predictors of injury mortality in infants. Their findings suggest that birth order may be an important predictor for infant injury death, with infants born to mothers with more than 2 other children at higher risk of injury mortality. Nathens et al. (2000) used Washington State birth, hospital discharge, and death data to study risk factors for unintentional injuries to children under age 6 years. Nathens found an association between the presence of older siblings and increased odds of injury in this older age group. Other important risk factors for childhood injury in this study include maternal age, maternal marital status, maternal education, prenatal care, insurance status, preterm birth, and gender.

Most studies of childhood injury group children into categories of < 1 year and 1–4 years. In a study using the California hospital discharge database, Agran et al. examined the mechanism of injury for infants using finer increments of time. These

investigators found that, in the first year of life, causes of infant injury change by each 3-month age grouping. In fact, infants in each age grouping had a different leading cause of injury: falls from height (0–2 months), battering (3–5 months), falls from furniture (6–8 months), and non-airway foreign body (9–11 months). While this study did not assess birth order, it does emphasize that infant injury is developmentally linked and may provide insight into associations of infant injury with birth order. Several other studies specifically examine the effect of birth order on childhood health, but do not focus on injury directly, rely on parent recall, or have small sample size.

2.2 Generalized Linear Models

Regression analysis and analysis of variance (ANOVA) are widely used in virtually all fields including medicine, business, marketing, logistics, agriculture, product development, economics, and more. These traditional linear models assume that the outcome being studied is normally distributed. The normality assumption is problematic when data do not fit this requirement, as is frequently the case. Nelder and Wedderburn (1972) first introduced "generalized linear models" to solve the problem of building a linear model for non-normally distributed outcomes.

Generalized linear models (GLM) include the aforementioned ordinary least squares methods (simple linear regression and ANOVA), as well as expansions of these traditional methods such as logistic regression, Poisson regression, and loglinear models to deal with categorical, count, and other non-normally distributed data (Neter et. al., 1996). By letting $\mathrm{Y} = [Y_1, Y_2, ...Y_n]$ represent the n independent observations from an outcome variable of interest, then generalized linear models may be characterized by two requirements (Rencher 2000):

1. The outcome variable, Y_i, has a density function from the exponential family (binomial, normal, Poisson, gamma, negative binomial, etc.).

2. A link function, g, of the expected value of the outcome variable is described by a linear function of predictors.

Requirement 2 is expressed as $g(E(Y_i)) = x_i'\beta$, which is equivalent to $E(Y_i) = g^{-1}(x_i'\beta)$. In this model, g must be a monotonic, differentiable function (McCullagh and Nelder 1989). The variance of the outcome variable also turns out to be specified as a function of the expected value because of the properties of exponential family distributions:

$$V_i = f(E(Y_i)) . \qquad (1)$$

Following notation similar to Liang and Zeger's (1986), the estimate of β is the solution to a set of k "quasi-score" differential equations for k covariates and N observations:

$$U_k(\beta) = \sum_{i=1}^{N} D_i V_i^{-1}\left(Y_i - E(Y_i)\right) = 0 , \qquad (2)$$

where $D_i = E(Y_i)/\beta$. If the model is specified correctly, then asymptotically $E[U_k(\beta)] = 0$ and $Cov[U_k(\beta)] = D_i' V_i^{-1} D_i$. Therefore, the function $U_k(\beta)$ behaves like the derivative of a log-likelihood; estimation is accomplished by generalized weighted least-squares, usually through an iterative process (McCullagh and Nelder 1989).

To summarize, a generalized linear model is a linear model for the transformed expected value of an outcome variable having a distribution from the exponential family of distributions. The generalized linear model only requires that a relationship between the expected value of the outcome variable and the explanatory variables and between the mean and variance of the response variable is specified. The primary attractiveness of the GLM is its allowance for linear and non-linear models under a single framework. It

is possible to fit models for data that are normal, gamma, Poisson, geometric, binomial, or from any distribution of the exponential family. Generalized linear models for binary data are a special subset of GLMs, as explained below.

2.3 Generalized Linear Models for Binary Data

Many response variables have only two possible outcomes. Examples include patients' contraction of bacterial infection during hospital stay (Yes, No), mortality from car crash (Died, Survived), decision to purchase a product (Buy, Not Buy), or choice of computer (Laptop, Desktop). In each of these scenarios, the response will always be one of two possible outcomes.

Agresti (1990) describes how to represent this class of outcomes statistically. For models with binary outcomes, the response variable is represented by Y. Because each response Y_i has 2 possible outcomes, denoted by 0 and 1, the Bernoulli distribution is an appropriate description of Y from the exponential family. In this case, $E(Y_i) = 1*Pr(Y_i = 1) + 0 * Pr(Y_i = 0) = Pr(Y_i = 1)$. Representing $E(Y_i) = Pr(Y_i = 1)$ by $\pi(x_i)$ demonstrates the outcome variable's dependence on the values of the explanatory variables, x_i.

In general, the goal of modeling a Bernoulli outcome variable is to describe the relationship of the explanatory variables to the probability of an event ($Y_i = 1$). A classical linear probability method, such as an ordinary least squares regression, may be applied to the analysis of binary data, but there are three primary problems with this method (Agresti 1990). First, the variance of an ordinary least squares model should be constant. The variance of a Bernoulli outcome variable is given by $V(Y_i) = \pi(x_i)[1 - \pi(x_i)]$. This shows the variance is not constant (that is, it depends on the explanatory variables' influence on the probability of an event); therefore, ordinary least squares estimators will

not be unbiased minimum variance estimators. Second, from a likelihood point of view, the ordinary least squares method is optimal for a normally distributed outcome variable. However, a Bernoulli outcome variable is not normally distributed, which implies that the sampling distribution for the ordinary least squares method is inappropriate. Third, because a probability is being modeled, $\pi(x_i)$ should be restricted to $\pi(x_i) > 0$ and $\pi(x_i) < 1$; the ordinary least squares method does not restrict $\pi(x_i)$ in this way.

A more appropriate model for binary outcome data would not require assumptions about normality or assumptions about constant variance and would simultaneously model the relationship between the probability of an event in Y and the values of the explanatory variables (Agresti 1990).

If the link function, g, is defined as $g(\pi(x_i)) = \ln [\pi(x_i)/(1- \pi(x_i)]$, then $\ln[\pi(x_i)/(1-\pi(x_i)] = x_i'\beta$ or $\pi(x_i) = \exp(x_i'\beta)/[1 + \exp(x_i'\beta)]$ (Dobson, 1990). By defining the link function this way, no normality assumptions or constant variance assumptions are made. For the single variable case, $\pi(x) = \dfrac{e^{(\alpha+\beta x)}}{1+e^{(\alpha+\beta x)}}$. Notice that when $\beta < 0$ and as $x \to 0$, $\pi(x) \to 0$ and when $\beta > 0$ and as $x \to \infty$, $\pi(x) \to 1$. Therefore, $\pi(x)$ has the appropriate restrictions on its value set for a single predictor variable (Agresti 1990). Dobson (1990) gives a more complete demonstration for multiple predictor variables.

The model defined here is called the logistic regression model, and the link function employed, g, is commonly referred to as the logit link function. In contrast to the ordinary linear regression model, the logistic regression model appropriately describes the relationship between the probability of an event in Y_i, expressed as $\pi(x_i)$, and the explanatory variables, expressed as x_i, because it does not require normality, does not assume constant variance, and restricts the outcome to (0,1).

The logistic regression model has useful properties relating to interpretation of results. The function, $\pi(x_i)/(1 - \pi(x_i)$, is usually referred to as the "odds." For example, if response "Yes" has odds of 3, then the response is 3 times as likely as response "No". When taking the ratio of the odds, this is called the "odds ratio." The odds ratio is a measure of association frequently used in medical and public health applications (Agresti 1990). The interpretation of the β parameters is straightforward and directly related to the odds ratio. The β parameters are interpreted as a multiplicative effect on the odds ratio (Stokes 1985). In the case of a Bernoulli explanatory variable, e^{β} is the estimate of the odds ratio of the outcome for one level of the explanatory variable compared to another level of the explanatory variable. For example, if the explanatory variable is sex and $e^{\beta} = 2$ for males compared to females, the parameter estimate would be interpreted as follows: "The odds of observing males with an event are increased two-fold when compared to females."

2.4 Generalized Estimating Equations

In many cases data are correlated. Common types of correlated data include repeated measures and clustered data. Observations on the same individual (repeated measures) or observations from individuals in the same cluster (clustered data) tend to exhibit correlation; thus, analysis without accounting for this relationship may result in poorly fitted models and will always result in incorrect estimates of the variances. For correlated data arising from repeated measurements where the measurements are assumed to be from a normal distribution, analysis methods have been investigated and fairly well-developed (Littell et al. 1996), although work is still needed in calculating degrees of freedom and in handling small sample sizes (Schaalje et al. 2002).

Although correlated data may be normally distributed, the normality assumption is problematic for many response variables of interest. This is especially true in the medical and health sciences where outcomes are frequently both correlated and event-oriented. In order to address this problem, Liang and Zeger (1986) proposed an extension of generalized linear models to the analysis of correlated data. This method is ideal for data that could otherwise be analyzed using a generalized linear model, except for the correlation among observations. This extension of generalized linear models introduces a class of estimating equations which allows for analysis of non-normally distributed and correlated data. This analysis technique, known as "Generalized Estimating Equations" (GEEs), provides a practical approach to the analysis of non-normal, correlated data.

2.4.1 An Overview of GEEs

The solution proposed by Liang and Zeger (1986) for expanding GLM's to correlated data is to specify a "working" correlation matrix incorporated into the variance term of equation (1). Let n_i be the number of observations within each cluster, and let j index the observations within a cluster: $j = [1, 2, \ldots n_i]$ for $i = [1, \ldots, K]$ clusters. Also let $R_i(\alpha)$ be an $n_i \times n_i$ correlation matrix for cluster i, where α is a vector which fully characterizes $R_i(\alpha)$. Equation (1) becomes a covariance matrix for the i-th cluster:

$$V_i = A_i^{1/2} R_i(\alpha) A_i^{1/2} / \phi, \qquad (3)$$

where the A_i are $n_i \times n_i$ diagonal matrices with $f(E(Y_{ij}))$ as the diagonal elements and ϕ is the scale parameter for exponential family distributions. Substitution of equation (3) into equation (2) gives us the general estimating equations:

$$U_k(\beta) = \sum_{i=1}^{K} D_i' V_i^{-1} \left(Y_i - E(Y_i)\right) = 0, \qquad (4)$$

where $D_i = [\partial(E(Y)/\partial(\beta)]$ is an $n_i \times k$ matrix, and $Y_i - E(Y_i)$ is of order $n_i \times 1$ for the i^{th} cluster. Notice that when $n_i = 1$, or the independence case, the GEE estimator reduces to a GLM. From this substitution, it is clear that GEEs are an extension of the GLM model. Therefore, the interpretation of the parameter estimates stems from the generalized linear model. For example, for a logistic regression model incorporating a correlation structure, the β parameter estimates are still interpreted in the same manner as in a logistic regression model; that is, as a multiplicative effect on the odds ratio.

2.4.2 Advantages of GEEs

As discussed and shown by Liang and Zeger (1986) and Lipsitz et al. (1994), GEEs have a number of advantages for the analysis of correlated data. First, GEEs offer reasonable statistical efficiency. Because the first two terms of equation (2) do not depend on Y_i, the score equations converge to 0, which implies that the score equations are consistent as long as $E[Y_i - E(Y_i)] = 0$. Additionally, when $E(Y_i)$ is correctly specified, GEE estimates of the parameters ($\hat{\beta}_r$) will also be consistent (Liang and Zeger 1986).

A second advantage to the GEE model is its allowance for a range of correlation structures within correlated groups (clusters). Three common specifications of the "working" correlation matrix $R_i(\alpha)$ include the exchangeable, autoregressive, and unstructured correlation structures. For the following examples, s and t are used to represent the rows and columns of R_i. Set $R_i(\alpha) = \rho,\ s \neq t$ and $R_i(\alpha) = 1,\ s = t$. This is an exchangeable correlation structure. This structure allows all of the Y_i to be related to each other in the same way across all observations in a cluster.

Setting $R_i(\alpha) = \rho^{|t-s|}$ generates an autoregressive correlation structure. The autoregressive correlation structure forces the correlation to be the same across observations and the correlation within an observation is an exponential function of the "distance" between observations. If $R_i(\alpha) = \alpha_{st}$, $s \neq t$, and $R_i(\alpha) = 1$, $s = t$, constraints on the correlation between observations within a cluster are removed and each correlation might be unique. There are a number of other correlation matrix specifications available, as described in detail by Fitzmaurice et al. (1993).

Finally, like GLMs, GEEs have applicability to a wide range of data including non-normal continuous, dichotomous, polychotomous, ordinal, and event-count response variables.

2.4.3 Inference from GEE Models

GEEs are a population-averaged (or marginal) approach to analyzing correlated data, which differs from the traditional conditional approaches to correlated data analysis. Neuhaus, Kalbfleisch, and Hauck (1991) as well as Hu et al. (1998) both provide a good discussion on the distinctions between these two approaches. They explain that conditional approaches, such as mixed model analysis, model the distribution of the response variable as a function of the predictor variables and a parameter specific to each cluster. The cluster-level parameter is estimated as a fixed-effect or as a random-effect. In marginal models, however, the population-averaged expectation of the dependent variable is modeled as a function of the predictor variables. There is no specific cluster-level parameter; instead, intracluster correlation is accounted for by specifying an appropriate covariance matrix to account for non-independence between observations. Diggle, Liang, and Zeger (1994, pg 131) explain this concept as follows, "Marginal

models, then, model the . . . average response over the sub-population that shares a common value of X." The distinction between the two types of models is important due to the difference in how the parameter estimates may be interpreted. A conditional model's parameter estimate represents the effect of a change in the predictor for the same individual, whereas a marginal model's parameter represents the average effect of a one-unit shift in the predictor across the entire population (Pendergast et al. 1996). Because GEEs do not explicitly model between-cluster variation, as conditional model approaches do, it is important to note that for GEEs the computational complexity is a function of the size of the largest cluster rather than the number of clusters. Therefore, when there are many small clusters, GEEs have a computational advantage over conditional models and are a source of reliable parameter estimates.

2.4.4 GEEs and Convergence, Goodness-of-Fit, and Missing Data

Under certain circumstances, GEEs may fail to converge. Generally, GEE convergence becomes more difficult as the number of clusters (sample size) decreases, as the number of correlation parameters being estimated increases, and as the size of intracluster correlations increases. Lipsitz et al. (1994) found that for *N* (clusters) = 15 and r = .60, 65% of convergence problems were traced to singularities in the variance-covariance matrix, and the other 35% were due to exploding estimates of β. Therefore, when using GEEs with few clusters or high intracluster correlations, there is a tradeoff between specifying a complex correlation matrix and computational manageability.

Another issue to consider when using GEE analysis is that goodness-of-fit statistics are problematic for GEE models. Because residuals from GEEs are correlated, they are not appropriately evaluated by many common goodness-of-fit procedures

(Chang 2000). There have been some recent developments in goodness-of-fit statistics for GEEs, however, they are all limited to the binary outcome variable case. These will be discussed in more detail in the final section.

In the context of GEE models, the issue of missing data needs special treatment. Little and Rubin (1987) and Sherman (2000) outline what are now the standard classifications for missing data: missing completely at random (MCAR) and missing at random (MAR). For models based on complete likelihoods, these types of random missing data are typically considered ignorable non-response mechanisms. Inferences are valid without explicitly modeling the mechanism for the missing data. For GEE models, this is not the case. Data which are MAR (past values of Y affect the probability of missingness), will not necessarily yield consistent estimates of β. Fitzmaurice, Laird, and Rotnitzky (1993) show that the extent of the bias in estimating β depends on several factors, including the extent of the missing data, accuracy of the model's specification, presence of time-varying covariates, and specification of the working correlation matrix.

2.5 Generalized Estimating Equations for Binary Data

GEE modeling has been developed primarily in the context of binary response variables. The following discussion focuses on two recent developments related to GEEs for binary data, goodness-of-fit statistics, and the alternating logistic regressions algorithm.

2.5.1 Goodness-of-Fit Statistics for GEEs

Obtaining goodness-of-fit statistics is problematic for GEEs because goodness-of-fit statistics are not based on the complete information maximum likelihood (conditional approach), but on a marginal model approach based on quasi-likelihood. Because GEES

are a marginal model, the widely used likelihood-ratio tests for testing goodness of fit of the model are not available for GEEs. Recently, several goodness-of-fit tests for GEEs have been developed, although they are all limited to the case of binary dependent variables. These alternative goodness-of-fit tests are asymptotically identical to likelihood ratio tests developed for ordinary least squares models.

Barnhart and Williamson (1998, pg 720) developed a test based on "partitioning the space of covariates into distinct regions and forming score statistics that are asymptotically distributed as chi-square random variables with the appropriate degrees of freedom." The primary drawback of this goodness of fit test is the necessity of partitioning the covariate space, which becomes cumbersome when many or continuous covariates are included in the model.

Horton et al. (1999) proposed an extension to the GEE context of the goodness-of-fit test for ordinary logistic regression developed by Hosmer and Lemeshow (1980). This statistic is constructed by estimating a model, generating predicted probabilities, dividing the data into G groups based on deciles of the predicted probabilities, defining G-1 indicator variables corresponding to the deciles, and then including the indicator variables in an additional model from which score or Wald statistics are derived. This statistic has an approximate chi-squared distribution when the model has been specified correctly, and a significant result indicates a lack of fit.

Additional approaches summarized by Zheng (2000, pg 1265) include tests based on reductions in entropy and deviance and on "the concordance correlation coefficient and the concordance index," which are indices of concordance between ordinal ranking

or predicted versus actual values. Future research is needed to compare these different approaches for goodness-of-fit of GEE models.

2.5.2 Alternating Logistic Regressions Algorithm

The Alternating Logistic Regression (ALR) algorithm was developed as a more computationally feasible option for obtaining parameter estimates than typical GEEs. Carey, Zeger, and Diggle (1993) introduced the ALR algorithm as an alternative method for taking the correlation between measurements for correlated binary data into account. Instead of using correlation between measurements to model association, as GEEs normally do, the log odds ratios may be used instead. The ALR algorithm iterates between a GEE step and a logistic regression step; the GEE step estimates regression coefficients, and the logistic regression step updates odds ratio parameters. When the ALR algorithm converges, it provides estimates of the mean and log odds ratios, as well as their regression parameters (standard errors and covariances) (SAS 2004). The primary reason for using the log odds ratios instead of the normal GEE is that GEEs become computationally unmanageable with large cluster sizes. When using the ALR algorithm to model association between pairs of responses, clusters of size n require inversion of matrices of order n^2 rather than n^4, making the ALR algorithm a more feasible approach if cluster sizes are large. The ALR algorithm estimates have also been shown to be reasonably efficient (Carey, Zeger, and Diggle 1993).

2.6 Conclusion

GEEs provide a fairly flexible and easily implemented method for analyzing correlated data. They are well-suited for data with many small clusters and provide a more computationally feasible alternative to full-likelihood approaches. Furthermore,

there is a wide range of software packages that offer GEE analysis as a built-in feature, providing estimates of odds ratios and other regression parameters that researchers are already familiar with from logistic regression of non-correlated data. Because of their utility, accessibility, and particular applicability to biostatistics and epidemiological studies, GEEs will continue to be an active area of development.

3 METHODS

This section provides detailed information about the data used for analysis, how the variables were structured and presented, and the analysis methods and statistical tests that were used to answer research questions.

3.1 Data

Data for analysis included Utah birth certificate, death certificate, and hospital emergency department datasets. The birth dataset contained birth certificate information for all live births occurring in Utah between the years 1999–2002. Information from the birth certificate dataset included maternal factors: age, race, education, marital status, alcohol and tobacco use during pregnancy, and sufficiency of prenatal care; as well as infant factors: gestational age, major birth abnormalities or anomalies, birth order, multiple-birth status (twins, triplets, etc.), and birth weight. The death certificate database provided information for all injury deaths that occurred within one year of the birth of an infant, 1999–2003. The emergency department database included information regarding whether or not an infant was seen in the emergency department and discharged or seen and admitted to the hospital for an injury during the first year of life. These three datasets were probabilistically linked using LinkSolv 7.0 (Strategic Matching, Inc.) to obtain complete infant, maternal, and emergency department information for all medically attended infant injuries. By definition, these data did not include information about injuries to infants who were born in Utah that were treated outside of Utah, or injuries that occurred to infants born outside of Utah who were treated in Utah.

Also excluded from the analysis were multiple-birth infants (twins, triplets, etc.), premature infants (completed gestational age ≤ 24 weeks), infants with a birth weight < 500 grams, and infants with a major birth defect noted on the birth certificate. Multiple-birth infants were excluded because they are difficult to distinguish from each other when conducting probabilistic linkage and will frequently be double-counted. Extremely premature births, very low birth weight infants, and infants with a major birth defect were excluded because these groups are likely to die or spend extended periods of time in the hospital after birth, thus changing their exposure time to risk of injury. Major birth defects noted on the birth certificate include spina bifida, anencephaly, hydrocephalus, microcephalus, renal agenesis, tracheo-esophageal fistula, esophageal atresia, gastroschisis, omphalocoeole, diaphragmatic hernia, chromosomal anomalies, multiple anomalies, birth injury requiring ventilation greater than 30 minutes, congenital infection, meconium aspiration requiring greater than 30 minutes ventilation, seizures, and congenital heart defects.

3.2 Risk Factors

Three groups of risk factors were defined for analysis: maternal demographics, maternal risk behaviors, and infant demographics.

Maternal demographics include level of education, age, race, and marital status. Maternal education is presented as "Less than High School," "Completed High School," or "Education Beyond High School." Maternal age is treated continuously. Maternal race is presented as "Non-Hispanic White," "Hispanic," or "Other Minority." Maternal marital status is "Not Married" versus "Married."

Maternal risk behaviors include adequacy of prenatal care, cigarette smoking

during pregnancy, and alcohol consumption during pregnancy. Adequacy of prenatal care is defined by the Kotelchuck index (Kotelchuck 1994), and presented as "Adequate" or "Inadequate." Cigarette smoking and alcohol consumption are both dichotomized into "No" versus "Yes." These variables are coded "Yes" if the mother self-reported smoking at least 1 cigarette or consuming at least 1 alcoholic drink during any trimester of her pregnancy. It is expected that these numbers are under-reported.

Infant demographics include sex, birth order, and prematurity (gestational age < 37 weeks). Sex is dichotomized into "Female" versus "Male." Birth order is 1^{st}, 2^{nd}, 3^{rd}, 4^{th} or greater birth order. The reference group is the firstborn child. Prematurity is defined as completed gestational age < 37 weeks and is dichotomized as "Premature" or "Not Premature."

Interactions between maternal age and several other covariates were considered in the model including birth order, maternal education, marital status, race, prenatal care, smoking behavior during pregnancy, and alcohol use during pregnancy. Additionally, interactions between maternal prenatal care and race, maternal prenatal care and maternal education, and birth order and maternal education were also considered. Finally, a quadratic term and a cubic term for maternal age were considered in the model.

3.3 Outcomes

Two outcome variables were defined for this study. The first outcome is an "injury event." An injury event is defined as an emergency department-attended injury, or a death resulting from injury. Patients who were seen in an emergency department (ED) or died during their first year of life due to injury were flagged in the dataset as an injury event. Some patients may have been seen in the emergency department more than

once during their first year of life; the injury outcome variable only indicates that at least one medically attended injury (or death) occurred. Injury events were identified using the World Health Organization's International Classification of Diseases, Ninth Revision, diagnosis codes and external cause of injury codes (Ecodes). Thus, diagnosis codes 800–999 were used to identify injuries, and wherever the Ecode was available, the injury was excluded if the injury was caused by medical procedures (E870–E879), adverse effects from treatment (E930–E949), legal intervention (E970–E979), or operations of war (E990–E999).

The second outcome is a "severe injury event." Patients who were admitted to the hospital or died during their first year of life due to an injury were flagged in the dataset as having a severe injury event and were flagged as no severe injury otherwise. Some patients may have been admitted to the hospital during their first year of life due to the same or different injury events multiple times; however, the severe injury outcome only indicates at least one hospital admission (or death) due to injury. Severe injury events were identified using death certificate records and ED records indicating whether or not an injured infant was admitted to the hospital.

3.4 Analysis

For each outcome variable, injury, and severe injury, the same analysis methods were used and the same procedure for each outcome was followed. First, the data were summarized by calculating means, medians, frequencies and percents. Next, a non-correlated subset of the complete dataset was created by randomly selecting one sibling from each family group. Then, logistic regression was used to analyze the non-correlated dataset. Due to controversy over which variable selection method is optimal

for model-building, multiple model building strategies including stepwise, forward and backward elimination methods were used to create three different models. The three models were compared for any differences in variable selection. No differences were found in the variables selected for each of the three models. Variables were kept in the logistic regression models with a $p < .05$. Last, the complete dataset was analyzed. Because the complete dataset contained siblings, observations were correlated within family groups. Therefore, the complete dataset was analyzed using generalized estimating equations (GEE), a method that incorporates correlation between observations.

The non-correlated was analyzed before the complete dataset because the variable selection and analysis of the complete dataset required computational resources that were unavailable. By using the non-correlated subset for variable selection and then further refining the model using the complete dataset, an appropriate model could be constructed without violating any assumptions of the analysis methods employed.

Injury incidence rate per 1,000 person-years was calculated for each subgroup of statistically significant interaction terms. These were calculated using the number of births in the subgroup as the denominator. Statistical significance was declared with a p-value of <0.05 in the adjusted final model. Odds ratios and 95% confidence intervals (CI) from the GEE analysis were used for presentation of the results. Because SAS automatically excludes any record without complete information for all of the covariates, the missing values population was tested for differences in the covariates compared with the analysis population using chi-squared tests of independence and t-tests. No statistically significant differences were found between the missing values population and the analysis population for each of the covariates. The non-correlated dataset was created

using PROC SURVEYMEANS, the non-correlated dataset was analyzed using PROC LOGISTIC, and the complete dataset was analyzed using PROC GENMOD (SAS 9.1.3 Cary, NC).

4 RESULTS

4.1 Population Description

Between the years 1999–2002, there were 195,070 live births in Utah. Excluded from the analysis were 8,395 infants with < 24 weeks gestational age, infants with a major birth defect, multiple-birth infants, and infants with birth weight < 500 grams. Also excluded are 507 infants whose deaths were due to non-injury causes.

Between the years 1999–2003, there were 8,553 infant injured in the state of Utah that met the eligibility criteria of this study. Of these infants, 637 were excluded because the infant's injury resulted from complications of medical care, birth injury, or other medical misadventures. Thus, the eligible study population was comprised of 185,531 infants; a total of 7,798 of these infants were injured.

4.2 Mother and Infant Demographics

Infant demographics are shown in Table 4.1. There was an incremental increase in the number of births each year with negligible difference in the proportion of males to females. There were 13,048 infants with < 37 weeks completed gestational age.

Table 4.1: Infant Demographics

		N	%
Year			
	1999	44767	24.1
	2000	46092	24.8
	2001	46779	25.2
	2002	47893	25.8
Sex			
	F	90324	48.7
	M	95206	51.3
	Missing	1	0.0
Birth Order			
	1st	66527	35.9
	2nd	54231	29.2
	3rd	33860	18.3
	4+	30315	16.3
	Missing	598	0.3
Premature (<37 wks)			
	Y	13048	7.0
	N	172483	93.0

Maternal demographics are displayed in Table 4.2. Mothers' median age was 26 years (25^{th}, 75^{th} quartiles: [22, 30]). Maternal education was high with over 52% of mothers having attained education beyond high school. The largest minority group in the state is Hispanic (13%). Most mothers (83%) were married.

Table 4.2: Maternal Demographics

		N	%
Age			
	<21	24429	13.2
	21–25	65870	35.5
	26–30	54132	29.2
	>30	41089	22.1
	Missing	11	0.0
Education			
	< high school	27955	15.1
	= high school	58519	31.5
	> high school	96621	52.1
	Missing	2436	1.3
Marital Status			
	Y	153595	82.8
	N	31935	17.2
	Missing	1	0.0
Race/Ethnicity			
	Non-Hispanic White	150337	81.0
	Hispanic	23614	12.7
	Other Minority	8277	4.5
	Missing	3,303	1.8

4.3 Maternal Risk Behaviors

Information about maternal risk behaviors available from the birth certificate data included extent of prenatal care, smoking during pregnancy, and consumption of alcohol

during pregnancy (Table 4.3). A majority of mothers (84%) received adequate prenatal care. Few mothers reported smoking (8%) or consuming alcohol during pregnancy (1%).

Table 4.3: Maternal Risk Behaviors

		N	%
Prenatal Care			
	Adequate	155396	83.8%
	Inadequate	24451	13.2%
	Missing	5684	3.1%
Consumed Tobacco During Pregnancy			
	Y	14386	7.8%
	N	170182	91.7%
	Missing	963	0.5%
Consumed Alcohol During Pregnancy			
	Y	2052	1.1%
	N	182443	98.3%
	Missing	1036	0.6%

4.4 GEE Analysis

4.4.1 All Infant Injuries

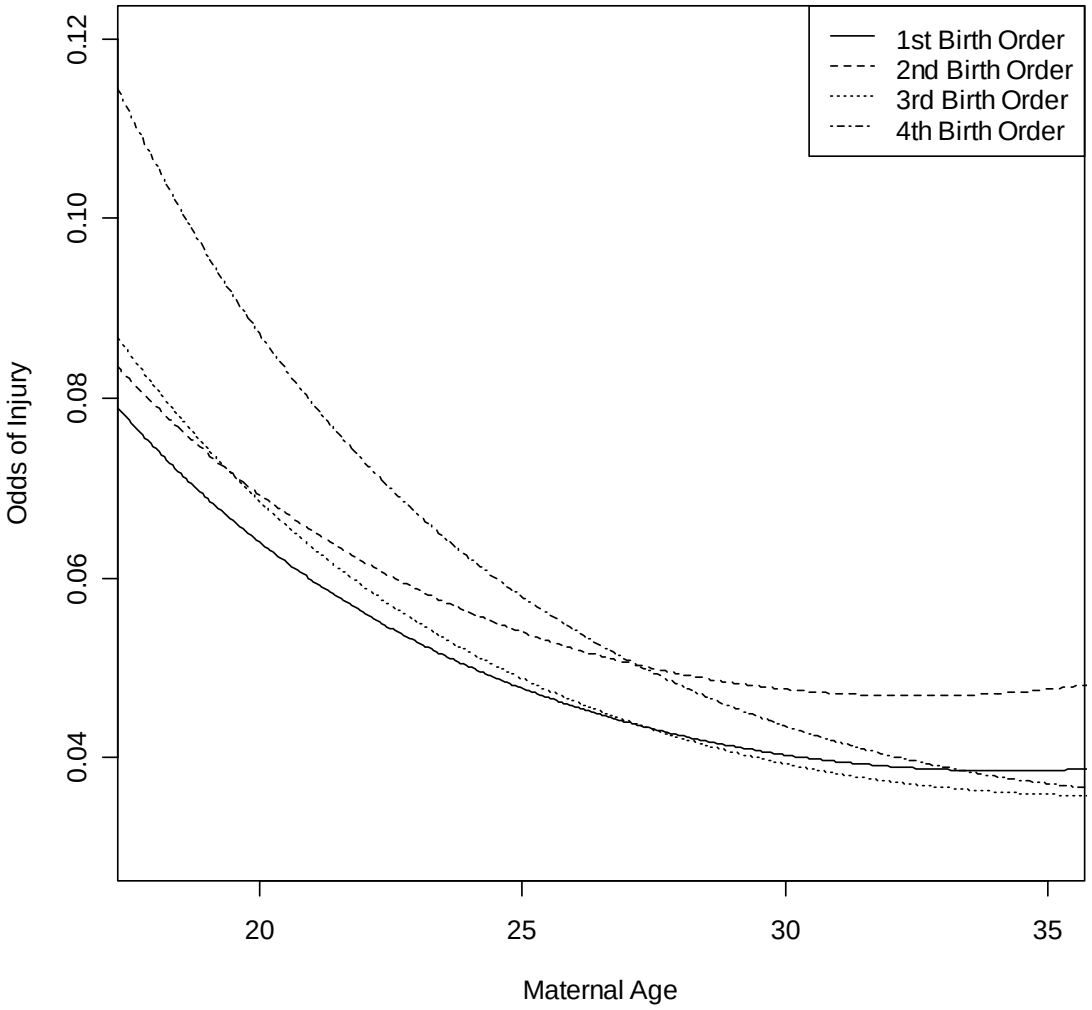
Complete model information and SAS output for all infant injuries are presented in Appendix A. Odds ratios and p-values for main effects are presented in Table 4.4.

Table 4.4: P-Values and Adjusted Odds Ratios for All Infant Injury

Risk Factor		Adjusted OR	Lower	Upper	P-Value
Year	2000	1.0	1.0	1.1	0.0614
Year	2001	1.0	1.0	1.1	0.1466
Year	2002	1.0	0.9	1.0	0.4884
Married	N	1.2	1.1	1.2	<.0001
Race/Ethnicity	Hispanic	0.8	0.8	0.9	<.0001
Race/Ethnicity	Other Minority	1.1	1.0	1.2	0.0048
Maternal Smoking	Y	1.2	1.1	1.2	<.0001

Birth order and year of birth were associated with infant injury. Sex and prematurity of the infant did not show an effect. There was a statistically significant interaction between birth order and maternal age (Figure 4.1). High birth order infants of young mothers have the highest odds of injury.

Figure 4.1: Odds of Injury by Birth Order and Maternal Age



Maternal characteristics associated with infant injury included marital status, race, age, and education. Infants' mothers who were unmarried had an increased odds of injury compared to infants of married mothers (AOR = 1.2, [95% CI, 1.1–1.2]).

Compared to non-Hispanic white mothers, infants of Hispanic mothers had decreased odds of injury (AOR = 0.8, [95% CI, 0.8–0.9]), while infants with a mother from other minority groups had increased odds of injury (AOR = 1.1, [95% CI, 1.0–1.2]). There was a statistically significant interaction term for maternal age and infant birth order. In general, as maternal age increased, the odds of infant injury decreased. Accounting for the interaction term, injury odds do not differ by infant birth order for mothers in their late twenties and older (Figure 4.1).

Table 4.5 provides the injury incidence rate per 1,000 person-years for infants by birth order and mother's age. The incidence rate of infant injuries decreased with mother's age, and infant birth order was associated with a higher incidence of infant injuries for younger mothers, but not for older mothers. The age-squared term for mother's age was also statistically significant. The adjusted odds of injury decreased until mothers reached their late twenties, and then injury odds remained constant.

Table 4.5: Incidence Rate per 1 000 person-years (95% CI): Birth Order by Maternal Age

Birth Order	Mother's Age			
	<21	21–25	26–30	>30
1	68 (64–71)	38 (36–40)	34 (31–38)	28 (23–32)
2	69 (62–76)	44 (41–47)	37 (34–40)	34 (30–38)
3	58 (41–75)	50 (46–55)	35 (32–38)	33 (30–37)
4+	104 (42–166)	60 (51–70)	42 (38–46)	33 (30–35)

Maternal education was associated with infant injury odds, and there was a statistically significant interaction between maternal education and prenatal care. Infants born to mothers with adequate prenatal care and a higher educational level had decreased odds of injury (0.8, [95% CI, [0.8–0.9]) while infants born to mothers with adequate prenatal care and a low educational level had higher odds of injury (1.1, [95% CI, [1.0–

1.2]). Education was less important when comparing mothers with inadequate prenatal care to those with adequate prenatal care (Table 4.6). That is, maternal educational level only showed an effect for infants whose mothers received adequate prenatal care.

Maternal risk behaviors associated with elevated injury odds included smoking during pregnancy and adequacy of prenatal care. Smoking during pregnancy was associated with increased infant injury odds (1.2, [95% CI, 1.2–1.2]), and, as previously discussed, there was a statistically significant interaction between adequacy of prenatal care and maternal education.

Table 4.6: Adjusted Odds Ratios for Prenatal Care by Maternal Education

	OR	Lower	Upper
>HS, adequate care	0.8	0.8	0.9
<HS, adequate care	1.1	1.0	1.2
*Compared to =HS, inadequate care			
>HS, inadequate care	0.9	0.8	1.0
<HS, inadequate care	1.0	0.9	1.1
*Compared to =HS, adequate care			

4.4.2 Severe Infant Injuries

Complete model information and SAS output for severe injuries are presented in Appendix B. Birth order was associated with severe infant injury (Table 4.7). Second- and third-born infants did not have increased odds of severe injury; however, fourth-or more-born infants had an increased odds of severe injury (AOR=2.2, [95% CI, 1.1–4.6]). The interaction term for mother's age by birth order was not statistically significant for severe injuries.

Three maternal factors were associated with severe infant injury: smoking, marital status, and age. Infants whose mothers reported smoking during pregnancy (AOR=1.8

[95% CI, 1.4–2.3]), were unmarried (AOR=1.4 [95% CI, 1.1–1.8]), or young had increased odds of severe injury. For each 1-year increase in mother's age, injury odds decreased by 8%. Prenatal care was not associated with severe infant injury.

Table 4.7: P-values and Adjusted Odds Ratios for Severe Infant Injury

Risk Factor		Adjusted OR	Lower	Upper	P-value
Birth Order	2	1.6	0.9	2.7	0.0969
Birth Order	3	0.9	0.5	1.9	0.8551
Birth Order	4+	2.2	1.1	4.6	0.0305
Smoke	Y	1.8	1.3	2.3	<.0001
Married	N	1.4	1.1	1.8	0.0167
Age		0.9	0.9	1.0	0.0217

5 DISCUSSION

5.1 Study Findings

This study identified several risk factors for infant injury. Specifically, birth order and maternal smoking behavior were associated with infant injury and severe infant injury. Maternal race/ethnicity was associated with infant injury, but not severe infant injury. Overall, this study ascertained two targeted groups well-suited for injury prevention efforts.

Similar to the findings of Nathens et al. (2005), Brenner et al. (1999), and Scholer et al. (1999), whose study settings were hospital admissions or deaths due to injury for slightly older children, this study found birth order was associated with infant injury. However, this study found that for all-cause ED-attended injuries, high birth order is most important among young mothers and not as important for older mothers. As maternal age increases, the effect of birth order diminishes. This is relevant because there is a relatively small group of infants born to young mothers with many other children in the household.

For severe injury (injury resulting in death or hospital admission), birth order did not become important until birth order was 4^{th} or greater. Interestingly, there was no interaction of birth order with maternal age for severe injuries as there was for all ED-attended injuries. Therefore, while high birth order was important for all injury in infants of young mothers, it is a risk factor for severe injuries in all families. Although the hypothesis had been that the risk of severe injury would increase as more children were added to the household, injury risk did not increase significantly until the fourth child entered the family. It may be that parent(s) gain experience with each new child, but

reach a threshold with the fourth child. It is possible that when there are many children in a household there is a large division of supervision from the parent(s), or an increase in supervision by older siblings resulting in higher risk of infant injury.

Maternal smoking behavior is associated with increased injury risk. This effect is largest for severe infant injury. Studies have demonstrated that persons who smoke are more likely to be injured from fires and to be in motor vehicle crashes (Sacks and Nelson 1994). Therefore, it may be that infants with mothers who smoke have increased risk of injury from fire and motor vehicle crashes. This is a noteworthy finding because the mothers in the population have a very low rate of smoking; therefore, a disproportionately large amount of infant injury risk is clustered within a relatively small group of infants.

Race/ethnicity showed an effect for ED-attended infant injuries. Specifically, the results of this study show a protective effect for infants born to Hispanic mothers. This finding is different from what has been found by other studies which have indicated that rates of pediatric injury are higher among Hispanics than among non-Hispanic whites in the United States (Agran et. al 1996). The difference between the results of this study and the results of others may be due to the fact that Hispanics are a relatively new population to Utah and may not have very good access to medical care, lack insurance, and tend to have lower household incomes. Anderson et. al (1998) found that poverty was associated with lower injury rates. Therefore, it may be that Hispanics in the Utah population do not have access to medical care because of poverty leading to lack of insurance and lack of access to medical care. The results of this study show a protective effect for mild-moderate injury (ED-attended injury), but no protective effect for severe

injuries (death or hospital admission). This implies that infants of Hispanic mothers in Utah may not have as much access to medical care for mild-moderate injuries as white and other racial/ethnic groups do, but utilize the ED for severe injuries as frequently as other racial/ethnic groups. This supports the idea that the protective effect observed in this study is probably due to an issue with access to medical care.

5.2 Limitations

Severely injured infants are frequently transferred to Intermountain Health Care (IHC) hospitals such as Primary Children's Hospital and LDS Hospital. Unfortunately, IHC hospitals do not provide infants' names to the Utah Department of Health (DOH). When patients are admitted from an emergency department to a hospital, the Utah Department of Health copies hospital information to the emergency department records. Without infants' names, the additional available information was not enough to link with high probability to the birth certificate file. This resulted in approximately 75% of severely injured infants not being identified as having had severe injury events. Assuming infants' names were missing at random, or similarly, that whether a severely injured infant was admitted as an inpatient to IHC or non-IHC hospitals was not related to other injury factors, not identifying 75% of the severe injury events implies that the results for severe injuries are conservative.

Although other studies have shown birth interval may be an important predictor for injury in households with multiple siblings, birth interval was not included as a predictor variable in the analysis because the data had 15% missing values for this field; including it would have caused the false loss of data for each observation without birth interval information. Birth order was important for infants of young mothers with many

other children in the household. This suggests that by looking at birth interval, the birth order effect may not be important after adjusting for birth interval because short birth intervals may account for the risk of injury in those families.

Overall, the model did not describe very much variation (~2%) in injury outcomes for infants, and many of the statistically significant results that were found are not practically meaningful for implementation in injury prevention. This suggests there may be other important infant injury risk factors not considered in the analysis.

5.3 Strengths

The study benefited from a large sample size, collected from statewide datasets. Using probabilistic linkage, this study examined information available from ED, birth and death data in a novel way. Additionally, using the GEE analysis was a new approach for studying infant injury because it allowed adjustment for family correlation structures.

5.4 Conclusions

A large proportion of infant injury risk is concentrated in a small percent of households, namely those with infants of 4^{th} or higher birth order with young mothers, and those with mothers who smoke. These two groups are more likely to have medically-attended injuries. Injury prevention efforts would be well-suited to these groups because it is a relatively small proportion of the population that has the largest amount of infant injury risk. Therefore, injury prevention efforts should be developed for parents and pediatricians to assist in reducing injury for these specific groups of infants.

5.5 Future Work

Imputation has been shown to perform well for up to 25% missing data in a single field, so future work might include imputation of birth interval, which currently has 15%

missing data, for inclusion in analysis to examine how this variable might also affect infant injury.

Another important avenue for future work is to investigate other options for identifying the severely injured infants more completely in the analysis dataset. Some options to consider are improving the probabilistic linkage by investigating other informative variables that might be available in both datasets, revising the linkage requirements, linking to the Utah state trauma registry, obtaining and linking to IHC hospital inpatient data, and/or imputing match status.

Although the model explained a small amount of total variation in infant injury, it did provide a targeted population for injury prevention. Future work should include investigation of additional factors that were not available in the datasets employed because additional explanatory variables may help further define infants at highest injury risk.

REFERENCES

Agran P.F., Winn D.G., Anderson C.L., Del Valle C.P. (1996), "Pediatric injury hospitalization in Hispanic children and non-Hispanic white children in southern California," *Arch Pediatr Adolesc Med.,*150(4),400–406.

Agran, P. F., Anderson, C., Winn, D., Trent, R., Walton-Haynes, L., Thayer, S. (2003), "Rates of pediatric injuries by 3-month intervals for children 0 to 3 years of age," *Pediatrics,* 111(6 Pt 1), 683–92.

Agresti, A. (1990), Categorical Data Analysis, Wiley-Interscience.

Anderson CL, Agran PF, Winn DG, Tran C (1998), "Demographic risk factors for injury among Hispanic and non-Hispanic white children: an ecologic analysis," *Injury Prevention*, 4(1), 33–38.

Barnhart, H. X. and J. M. Williamson (1998), "Goodness-of-fit Tests for GEE Modeling with Binary Responses," *Biometrics,* 54, 720–729.

Bijur, P. E., Golding, J., Kurzon, M. (1988), "Childhood accidents, family size and birth order," *Soc Sci Med,* 26(8), 839–43.

Brenner, R. A., Overpeck, M. D.,Trumble, A. C., DerSimonian, R., Berendes, H. (1999), "Deaths attributable to injuries in infants, United States, 1983–1991," *Pediatrics,* 103(5 Pt 1), 968–74.

Carey, V., S. L. Zeger, Diggle, P. (1993), "Modelling Multivariate Binary Data with Alternating Logistic Regressions," *Biometrika,* 80, 517–526.

CDC (2002). WISQARS (Web-based Injury Statistics Query and Reporting System), National Center for Injury Prevention and Control-Center for Disease Control (CDC).

Chang, Y. C. (2000), "Residual Analysis of the Generalized Linear Models for Longitudinal Data," *Statistics in Medicine,* 19, 1277–1293.

Danseco, E. R., Miller, T. R., Spicer, R. S. (2000), "Incidence and costs of 1987–1994 childhood injuries: demographic breakdowns," *Pediatrics,* 105(2), 27.

Diggle, P. J., K. Y. Liang, and Zeger, S.L. (1994), Analysis of Longitudinal Data, Oxford, Clarendon Press.

Dobson, A. (1990), An Introduction to Generalized Linear Models, London, Chapman and Hall.

Fitzmaurice, G. M., Laird, N.M., Rotnitzky, A.G. (1993), "Regression Models for Discrete Longitudinal Responses," *Statistical Science,* 8, 284–309.

Horton, N.J., Bebchuk, J.D., Jones, C.L., Lipsitz, S.H., Catalano, P.J., Zahner, G.E., Fitzmaurice, G.M. (1999), "Goodness-of-Fit for GEE: An Example with Mental Health Service Utilization," *Statistics in Medicine,* 18, 213–222.

Hosmer, D. W. and S. Lemeshow (1980), "Goodness of Fit Tests for the Multiple Logistic Regression Model," *Communications in Statistics Part A: Theory and Methods,* 9,1043–1069.

Hu, F.B., Goldberg, J., Hedeker D., Flay B.R., and Pentz M.A. (1998), "Comparison of Population-Averaged and Subject-Specific Approaches for Analyzing Repeated Binary Outcomes," *American Journal of Epidemiology,* 147, 694–703.

Kaplan, B. A., C. G. Mascie-Taylor, Boldsen, J. (1992), "Birth order and health status in a British national sample," *J Biosoc Sci,* 24(1), 25–33.

Kotelchuck, M. (1994), "An evaluation of the Kessner Adequacy of Prenatal Care Index and the proposed Adequacy of Prenatal Care Utilization Index," *American Journal of Public Health,* 84, 1414–1420.

Liang, K. Y. and S. L. Zeger (1986), "Longitudinal Data Analysis Using Generalized Linear Models," *Biometrika,* 73, 13–22.

Lipsitz, S. H., Fitzmaurice, G.M., Orav, E.J., Laird, N.M. (1994), "Performance of Generalized Estimating Equations in Practical Situations," *Biometrics* 50: 270–278.

Littell, R. C., Milliken, G.A., Stroup, W.W., Wolfinger, R.D. (1996), SAS System for Mixed Models, Cary NC, SAS Institute Inc.

Little, J. A. R., Rubin, D.B. (1987), Statistical Analysis with Missing Data, New York, Wiley.

McCaig LF, B. C. (2001), "National Hospital Ambulatory Medical Care Survey: 1999 Emergency Department Summary," Department of Health and Human Service. Centers for Disease Control and Prevention, National Center for Health Statistics.

McCullagh, P. and Nelder , J.A. (1989), Generalized Linear Models (Second Edition), London, Chapman and Hall.

Nathens, A. B., Neff, M.J., Goss, C. H., Maier, R. V., Rivara, F. P. (2000), "Effect of an older sibling and birth interval on the risk of childhood injury," *Injuryj Prevention*, 6(3), 219–22.

Nelder, J. A. and R. W. M. Wedderburn (1972). "Generalized Linear Models." *Journal of the Royal Statistical Society. Series A (General),* 135(3), 370–384.

Neter, J., M. H. Kutner, Nachtsheim, C.J., Wasserman, W. (1996), Applied Linear Statistical Models (Fourth Edition), Chicago, IRWIN.

Neuhaus, J. M., J. D. Kalbfleisch, Hauck, W.W. (1991), "A Comparison of Cluster-Specific and Population-Averaged Approaches for Analyzing Correlated Binary Data," *International Statistical Review,* 59, 25–35.

Pendergast, J., Gange, S.J., Newton, MA, Lindstrom, MP, Palta, M., Fisher, M.R. (1996), "A Survey of Methods for Analyzing Clustered Binary Response Data," *International Statistical Review,* 64, 89–118.

Rencher, A. C. (2000), Linear Models in Statistics, Wiley-Interscience.

Sacks, J.J., Nelson, D.E. (1994), "Smoking and injuries: an overview," *Preventive Medicine*, 23(4), 515–520.

SAS (2004), SAS OnlineDoc® 9.1.3. Cary, NC, SAS Institute Inc.

Schaalje, G.B., McBride, J.J. and Fellingham, G.W. (2002), "Adequacy of approximations to distributions of test statistics in complex mixed linear models," *Journal of Agricultural, Biological and Environmental Statistics*, 7, 512–524.

Scholer, S. J., Mitchel, E. F., Jr., Ray, W. A. (1997), "Predictors of injury mortality in early childhood," *Pediatrics,* 100(3 Pt 1), 342–347.

Sherman, R. P. (2000), "Tests of Certain Types of Ignorable Nonresponse in Surveys Subject to Item Nonresponse of Attrition," *American Journal of Political Science,* 44, 356–368.

Stokes, M. E., Davis, C.S., and Koch, G.G. (1995), Categorical Data Analysis Using the SAS System, Cary NC, SAS Institute Inc.

Sutton, P., Mathews, TJ (2004), Trends in Characteristics of Births by State: United States, 1990, 1995, and 2000–2002, *National Vital Statistics Reports*. D. o. V. Statistics, Centers for Disease Control.

van den Bosch, W. J., F. J. Huygen, van den Hoogen, H. J., van Weel, C. (1992), "Morbidity in early childhood, sex differences, birth order and social class," *Scand J Prim Health Care,* 10(2), 118–123.

Zheng, B. (2000), "Summarizing the Goodness of Fit of Generalized Linear Models for Longitudinal Data," *Statistics in Medicine,* 19, 1265–1275.

Weiss H, M. L., Forjuoh S, Kinnane J. (1997), Child and Adolescent Emergency Department Visit Databook, Pittsburgh, Pennsylvania, Center for Violence and Injury Control, Allegheny University of the Health Sciences.

APPENDIX A: ALL INJURIES, SAS CODE AND OUTPUT

```
ODS RTF FILE="P:\Users\hvanduker\Infant Injury First-time Mother\ANALYSIS\SAS Code and
Tables\TABLES\GEE.RTF";
ODS SELECT ALL;
PROC GENMOD DATA=BIRTH.dat2analysis DESC;
        format  BIRTHORDER SEX YEAR RACE3 MATOBACCO care3 MAEDU2 MAAGE MAMARRIED SEX;
        CLASS   MOTHERID
                BIRTHORDER(REF=FIRST) SEX (ref=first) NYEAR(REF=FIRST)
                MAMARRIED(ref=last) MAEDU2(ref=first) race3(ref=last)
                MATOBACCO(ref=first) CARE3(ref=LAST);
        MODEL   INJURY =        BIRTHORDER NYEAR
                MAMARRIED MAAGE MAAGE2 MAEDU2 RACE3
                MATOBACCO CARE3
                MAAGE*BIRTHORDER CARE3*MAEDU2
        /DIST=BINOMIAL LINK=LOGIT TYPE3;
        REPEATED SUBJECT=MOTHERID/TYPE=EXCH;
RUN;
ODS RTF CLOSE;
```

Score Statistics For Type 3 GEE Analysis			
Source	DF	Chi-Square	Pr > ChiSq
Birth Order	3	15.17	0.0017
Year	3	9.31	0.0254
Maternal Marital Status	1	60.17	<.0001
Maternal Age	1	52.31	<.0001
Maternal Age^2	1	32.22	<.0001
Maternal Education	2	17.58	0.0002
Race/Ethnicity	2	36.91	<.0001
Maternal Smoking	1	64.78	<.0001
Prenatal Care	1	17.08	<.0001
Age*Birth Order	3	13.72	0.0033
Education*Prenatal Care	2	15.08	0.0005

Analysis Of GEE Parameter Estimates						
Empirical Standard Error Estimates						
Parameter			Estimate	95% Confidence Limits		Pr > |Z|
Intercept			-0.3966	-0.943	0.1495	0.1546
Birth Order	2		-0.0949	-0.419	0.2293	0.5662
Birth Order	3		0.2569	-0.171	0.6844	0.239
Birth Order	4		0.7728	0.2329	1.3127	0.005
Year	2000		0.0382	-0.002	0.0782	0.0614
Year	2001		0.0295	-0.01	0.0694	0.1466
Year	2002		-0.0144	-0.055	0.0264	0.4884
Marital Status	N		0.1415	0.1069	0.1761	<.0001
Maternal Age			-0.1712	-0.213	-0.129	<.0001
Maternal Age^2			0.0025	0.0018	0.0033	<.0001
Maternal Education	< HS		0.0249	-0.035	0.0848	0.4149
Maternal Education	>HS		-0.1061	-0.164	-0.048	0.0003
Race/Ethnicity	Hispanic		-0.1756	-0.239	-0.113	<.0001
Race/Ethnicity	Other		0.109	0.0333	0.1847	0.0048
Maternal Smoking	Y		0.1759	0.1364	0.2154	<.0001
Prenatal Care	Adequate		0.0723	0.0367	0.108	<.0001
Age*Birth Order	2		0.0087	-0.005	0.022	0.1999
Age*Birth Order	3		-0.0094	-0.025	0.0064	0.2432
Age*Birth Order	4		-0.0232	-0.041	-0.005	0.0125
Education*Prenatal Care	<HS	Adequate	0.0958	0.0459	0.1456	0.0002
Education*Prenatal Care	>HS	Adequate	-0.072	-0.122	-0.022	0.0048

APPENDIX B: SEVERE INJURIES, SAS CODE AND OUTPUT

```
ODS RTF FILE="P:\Users\hvanduker\Infant Injury First-time Mother\ANALYSIS\MODEL BUILDING
SAS OUTPUT FILES\SEVERE\GEE_REMOVE MAAGEBYMAEDU2 MAEDU2.RTF";
ODS SELECT ALL;
PROC GENMOD DATA=BIRTH.dat2analysis DESC;
        format  BIRTHORDER MATOBACCO MAMARRIED MAEDU2;
        CLASS    MOTHERID
                BIRTHORDER(REF=FIRST)
                MATOBACCO(ref=first) MAMARRIED(ref=last) MAEDU2 (REF=FIRST);
        MODEL   SEVERE = BIRTHORDER MATOBACCO MAMARRIED MAAGE
        /DIST=BINOMIAL LINK=LOGIT TYPE3;
        REPEATED SUBJECT=MOTHERID/TYPE=EXCH;
RUN;
ODS RTF CLOSE;
```

Score Statistics For Type 3 GEE Analysis			
Source	DF	Chi-Square	Pr > ChiSq
Birth Order	3	8.26	0.0410
Maternal Smoking	1	10.08	0.0015
Marital Status	1	4.93	0.0264
Maternal Age	1	6.34	0.0118

Analysis Of GEE Parameter Estimates							
Empirical Standard Error Estimates							
Parameter		Estimate	Standard Error	95% Confidence Limits		Z	Pr > |Z|
Intercept		-5.7077	0.7016	-7.0829	-4.3325	-8.13	<.0001
Birth Order	2	0.4543	0.2736	-0.0820	0.9906	1.66	0.0969
Birth Order	3	-0.0638	0.3492	-0.7481	0.6206	-0.18	0.8551
Birth Order	4	0.7950	0.3675	0.0747	1.5153	2.16	0.0305
Maternal Smoking	Y	0.5705	0.1384	0.2992	0.8418	4.12	<.0001
Marital Status	N	0.3304	0.1381	0.0598	0.6010	2.39	0.0167
Maternal Age		-0.0735	0.0320	-0.1362	-0.0107	-2.30	0.0217