Modeling the Type and Timing of Consecutive Events: Application to Predicting Preterm Birth in Repeated Pregnancies

Joanna H. Shih
Biometric Research Branch, National Cancer Institute,
9609 Medical Center Drive, Rm 5W124, Bethesda MD, 20892
email: jshih@mail.nih.gov

Paul S. Albert
Biostatistics and Bioinformatics Branch
Division of Intramural Population Health Research
Eunice Kennedy Shriver National Institute of Child Health and Human Development,
6100 Executive Boulevard, Rockville MD, 20852

Pauline Mendola and S. Katherine Laughon
Epidemiology Branch
Division of Intramural Population Health ResearchResearch
Eunice Kennedy Shriver National Institute of Child Health and Human Development,
6100 Executive Boulevard, Rockville MD, 20852

November 21, 2014

Summary

Predicting the occurrence and timing of adverse pregnancy events such as preterm birth is an important analytical challenge in obstetrical practice. Developing statistical approaches that can be used to assess the risk and timing of these adverse events will provide clinicians with tools for individualized risk assessment that account for a woman's prior pregnancy history. Often adverse pregnancy outcomes are subject to competing events; for example, interest may focus on the occurrence of preeclampsia-related preterm birth, where preterm birth for other reasons may serve as a competing event. In this article, we propose modeling the type and timing of adverse outcomes in repeated pregnancies. We formulate a joint model, where types of adverse outcomes across repeated pregnancies are modeled using a polychotomous logistic regression model with random effects, and gestational ages at delivery are modeled conditional on the types of adverse outcomes. The correlation between gestational ages conditional on the adverse pregnancies is modeled by the semi-parametric normal copula function. We present a two-stage estimation method and develop the asymptotic theory for the proposed estimators. The proposed model and estimation procedure are applied to the NICHD Consecutive Pregnancies Study data and evaluated by simulations.

Keywords: adverse pregnancy outcome, preterm birth, preeclampsia, repeated pregnancies, normal copula, polychotomous random effects logistic model

1. INTRODUCTION

Adverse pregnancy outcomes affect the health of both the mother and the fetus. Women tend to repeat adverse outcomes in subsequent pregnancies, but our ability to predict recurrence for an individual woman remains limited (Louis et al., 2006). Developing individualized assessment of risk in consecutive pregnancies could result in risk stratification that both improves the predictive value of early pregnancy screening tests and facilitates development of individualized management plans. The NICHD (National Institute of Child Health and Human Development) Consecutive Pregnancies Study was designed to assess the association of adverse pregnancies and their timings during gestation across repeated pregnancies. Specifically, the primary objective of this study was to estimate the incidence of adverse pregnancy outcomes, to identify risk factors associated with this incidence, and to estimate individual risk incidence curves as a function of gestational age that account for both risk factors and prior pregnancy history.

The study collected retrospective data on 114,630 pregnancies from 50,166 women delivering > 20 weeks of gestation from 20 hospitals in the state of Utah from 2002 to 2010. The data set captured all consecutive births (> 20 weeks of gestation) within this nine year time period.

One of the current challenges in evaluating subsequent risk of adverse pregnancy outcomes is that prior history of one outcome increases the risk not only of that particular outcome but a number of different outcomes in a subsequent pregnancy, potentially due to similar biologic pathways. For example, history of medically indicated preterm birth is associated not only with increased risk of subsequent medically indicated preterm birth but also spontaneous preterm birth (Ananth et al., 2006). Therefore, at each pregnancy, women may be at risk of multiple types of adverse pregnancy outcomes where the occurrence of one event precludes the occurrence of another. In this article, we study two different indications for preterm birth < 37 weeks of gestation that are types of competing adverse pregnancy outcomes: (i) preterm birth complicated by preeclampsia designated as preeclampsia-related (PR) preterm birth and (ii) the remaining preterm births complicated by other medications or spontaneous labor, designated as preeclampsia-unrelated (PU) preterm birth. Preeclampsia is a syndrome in which a pregnant women develops new-onset high blood pressure and protein in the urine, after 20 weeks gestational age. The only treatment for preeclampsia is delivery, which may be required preterm (so called PR preterm birth). PR preterm birth may be associated with different risk factors from other preterm births, and studying these associations may be important for managing pregnant women who have had prior PR preterm births. Recent research in obstetrics has focused on predicting cause-specific preterm birth from prior pregnancy history and other etiological factors (Laughon et al., 2013). However, due to the lack of a statistical approach, predicting the timing as well as the occurrence has not been studied. This paper proposes a competing risk formulation for modeling the dual outcome of adverse pregnancy occurrence and gestational age. Specifically, our focus will be on modeling the type and gestational age of preterm birth, where PR and PU preterm birth are two competing events. The adverse preterm births and their timing are different from the conventional competing risk data in two aspects. First, the risk of a preterm birth is capped at 37 weeks after which the birth is classified as a term birth. In other words, the risk of having a preterm birth after 37 weeks is zero, and as such, term birth is not a competing event. Second, gestational age is observed for each pregnancy and thus not subject to censoring. Due to censoring, much of the work in competing risk literature (Kalbfleisch and Prentice (2002), and references therein) has been focused on modeling the cause-specific hazard function, even though it is a difficult to interpret quantity. In the current application, it is more desirable to correlate the gestational age of preterm birth with risk factors directly as compared with modeling the cause-specific hazard. Also, the investigators are interested in assessing the event (pre-term birth) process alone as well as together with the timing of the consecutive events. To this end, we develop a joint model to establish the inter-relationship between the two types of preterm births and their gestational ages at delivery in repeated pregnancies. In this formulation, types of preterm birth are modeled with a polychotomous logistic model using random effects, and repeated gestational ages at delivery are modeled conditional on the types of preterm births. The correlation between types of preterm birth in repeated pregnancies is induced by the type-specific random effects, and the correlation between gestational ages conditional on the types of preterm birth is modeled by the normal copula function. With the proposed joint model, the aforementioned quantities of interest such as the incidence of recurrent adverse pregnancy outcome and its relationship to the occurrence and timing of adverse outcomes on previous pregnancy are readily derived.

The article is organized as follows. In Section 2, we give an overview of the NICHD Consecutive Pregnancies study. In Section 3, we describe the proposed model and estimation method and present the asymptotic theory. We present the NICHD Consecutive Pregnancies Study analysis results in Section 4, and evaluate the performance of the proposed estimators with simulations in Section 5. We conclude the paper with a brief discussion in Section 6.

2. NICHD CONSECUTIVE PREGNANCY STUDY

An important objective of the study was to characterize the association in adverse pregnancy outcomes and their timing across consecutive pregnancies. As such, pregnancy, labor and delivery medical records for women in the state of Utah who had at least 2 pregnancies during 2002-2010 were retrospectively retrieved. Gestational age at delivery and type of pregnancy outcome were collected at each pregnancy during the 9 years of study period. For the purpose of analyses, we consider a woman to have preeclampsia during her pregnancy if she has the more serious conditions of eclampsia or superimposed preeclampsia. Eclampsia is defined as preeclampsia followed by seizure and superimposed preeclampsia is preeclampsia among women with chronic hypertension. A large number of clinical variables and pregnancy history prior to 2002 were collected as well.

Information on a total of 114,630 pregnancies from 51,066 women were recorded in the study. Table 1 tabulates the number of pregnancies and preterm births contributed by the women. Of the 114,630 pregnancies, 9,552 were preterm births delivered by 7,794 women. Figure 1(a) displays the histogram of gestational age of preterm birth. As the gestational age of preterm birth is capped at 37 weeks, the distribution is skewed to the left. Of these women, 1,571 (20.2%) had at least two preterm births, which together accounted for 3,329 (34.9%) preterm births, suggesting preterm births tend to recur. Of the 9,552 preterm births, 1043 (10.9%)were PR. Table 2 tabulates the 1,571 women with multiple preterm births by PR preterm birth and PU preterm birth. For example, 143 women had one PR and one PU preterm birth. As the two types of preterm birth occurred from the same woman, we explored whether the two types of preterm birth were correlated or not. Figure 1(b) displays the normal rank score of gestational age of the first pregnancy versus that of the second pregnancy among women who delivered preterm births in both pregnancies, where the squares and solid lowess smooth correspond to PR preterm births in both pregnancies, cross and dotted lowess smooth to PU preterm births in both pregnancies, and triangles and dotdashed lowess smooth to one of each type. It appears that the strength of correlation between gestational ages of preterm births might depend on the types of preterm births. In addition, a quick assessment of the effect of adverse pregnancy outcome on the subsequent pregnancy can also be illustrated by the conditional cumulative incidence. Figure 2(a) displays the cumulative incidence curve of the second pregnancy, where solid line corresponds to the marginal cumulative incidence of PR preterm birth, and the dashed and dotted lines correspond to the conditional counter part given the women had PR preterm birth and PU preterm birth at first pregnancy, respectively. Figure 2(b) displays the similar cumulative incidence plot for PU preterm birth. In both cases, there is a larger increase in the incidence of the same type of preterm birth than different types. For example, the marginal chance of having a PR preterm delivery by week 32 at the second pregnancy was 0.14%, but such a chance was increased to 0.21% and 4.4% if the woman had PU and PR preterm birth at first pregnancy, respectively.

[Figure 1 about here.]

[Figure 2 about here.]

The above brief summary demonstrates several features of the consecutive pregnancy data including 1) the correlation of same type as well as different type of repeated pregnancy outcomes; 2) the distribution of gestational age of preterm birth being skewed; 3) the correlation of gestational ages of multiple pregnancies depending upon the types of pregnancy outcomes; and 4) the association between multiple pregnancy outcomes and gestational ages. Developing personalized cumulative incidence curves is an important goal for clinical management of pregnant women. In the next section, we present a statistical model that will allow us to incorporate clinical history at each consecutive pregnancy in order to develop individualized risk prediction.

[Table 1 about here.]

[Table 2 about here.]

3. METHOD

3.1 Model

Consider a sample of n women of whom the pregnancy history data were collected. For the ith women, let $\mathbf{Y}_i = (Y_{i1}, \cdots, Y_{ik_i})$ and $\mathbf{T}_i = (T_{i1}, \cdots, T_{ik_i})$ denote multinomial pregnancy outcomes and gestational ages of k_i pregnancies, where $Y_{ik} = 0, 1, 2$ corresponds to term, PR preterm, and PU preterm birth, respectively. Let $\mathbf{Z}_i^Y = (\mathbf{Z}_{i1}^Y, \cdots, \mathbf{Z}_{ik_i}^Y)^t$ and $\mathbf{Z}_i^T = (\mathbf{Z}_{i1}^T, \cdots, \mathbf{Z}_{ik_i}^T)^t$ be $k_i \times p_1$ and $k_i \times p_2$ covariate matrices associated with $\mathbf{Y}_i$ and $\mathbf{T}_i$, respectively. The two sets of covariates may overlap. We model the joint distribution of $(\mathbf{Y}_i, \mathbf{T}_i)$ via $\mathbf{Y}_i$ and $\mathbf{T}_i \mid \mathbf{Y}_i$. The joint distribution of $\mathbf{Y}_i$ is specified through the polychotomous logistic regression with random effects model

$$p_{ijc}(b_{ic}) = p(Y_{ij} = c \mid b_{ic}) = \frac{\exp(\alpha_{0c} + \boldsymbol{\alpha}_c' \mathbf{Z}_{ij}^Y + b_{ic})}{1 + \sum_{c=1}^{2} \exp(\alpha_{0c} + \boldsymbol{\alpha}_c' \mathbf{Z}_{ij}^Y + b_{ic})}, \tag{1}$$

where $b_{ic}, c = 1, 2$ are the random intercepts indexing person-specific sensitivity to PR preterm birth (c=1) and PU preterm birth (c=2), respectively. It is assumed that $\mathbf{b}_i = (b_{i1}, b_{i2})$ follows a bivariate zero-mean normal distribution with variances σ_c^2 and correlation coefficient ν. Conditional on $\mathbf{b}_i = (b_{i1}, b_{i2})$, $(Y_{i1}, \cdots, Y_{ik_i})$ are assumed to be independent with the conditional probability function given by

$$p(\mathbf{y}_i \mid \mathbf{b}_i) = p(Y_{i1} = y_1, \cdots, Y_{ik} = y_k \mid \mathbf{b}_i) = \prod_{j=1}^{k_i} \prod_{c=0}^{2} p_{ijc}(b_{ic})^{I(y_{ij}=c)}, \tag{2}$$

where $p_{ij0} = 1 - \sum_{c=1}^{2} p_{ijc}, i = 1, \cdots, n$. The correlated random intercepts are used to induce correlation between multiple adverse pregnancies of the same type as well as of different types in repeated pregnancies. For example, if ν is positive, then a woman with a large value of b_1 has a higher chance of having PR preterm birth, which in turn increases the chance of having PU preterm birth in subsequent pregnancies. The joint distribution function of $\mathbf{Y}_i$ is obtained by integrating $\mathbf{b}$ in (2) and equals

$$p(\mathbf{y}_i) = \int p(Y_{i1} = y_1, \cdots, Y_{ik} = y_k \mid \mathbf{b}) g(\mathbf{b})\, d\mathbf{b},$$

where $g(.)$ is the density function of the bivariate normal distribution with mean $\mathbf{0}$ and variance-covariance matrix $\begin{pmatrix} \sigma_1^2 & \sigma_1\sigma_2\nu \\ \sigma_1\sigma_2\nu & \sigma_2^2 \end{pmatrix}$. Next, we model the joint distribution of $\mathbf{T}_i = (T_{i1}, \cdots, T_{ki})$ conditional on $\mathbf{Y}_i$. Since our interest is on inference related to PR and PU preterm birth, the distribution of gestational age beyond 37 weeks is not of interest and is not modelled. Correspondingly, modeling the joint distribution of gestational ages in the first 37 weeks given the pregnancy outcomes is equivalent to modeling the joint distribution of gestational ages among the pregnancies with $Y_{ij}'s > 0$. Since the conditional distribution of T_{ij} given $Y_{ij} > 0$ is skewed as seen in Figure 1(a), no distributional assumption is made for T_{ij}. Rather, we assume that T_{ij} (or a suitable transformation of T_{ij}) given $Y_{ij} > 0$ follows the linear model

$$T_{ij} = \beta_0 + \boldsymbol{\beta}'\mathbf{Z}_{ij}^T + \theta I(Y_{ij} = 1) + \epsilon_{ij}, j = 1, \cdots, l_i, \tag{3}$$

where $l_i = \sum_{j=1}^{k_i} I(Y_j > 0)$, is the number of preterm births for woman i , and ϵ_{ij} is the error term of which the distribution function is left unspecified. For model parsimony, we assumed that gestational age depends on the type of preterm birth (Y) of the same pregnancy only, not of different pregnancies. This is the reproducibility assumption often used in modeling the marginal distribution in multivariate outcomes analysis (Whittemore, 1995).

The correlation of repeated gestational ages of preterm births is specified by the joint distribution of ϵ_{ij}'s which is assumed to follow a normal copula model given by

$$(G(\epsilon_{i1}), \cdots, G(\epsilon_{il_i})^T) \sim \mathbf{\Phi}_\rho \tag{4}$$

for unspecified monotone increasing transformation G, where $\mathbf{\Phi}$ is the standard multivariate normal with common correlation coefficient ρ. Define the marginal distribution of ϵ_{ij} by $F(e) = \Pr(\epsilon_{ij} \leq e)$. As implied by the normal copula model (4), $G(.) = \Phi^{-1}\{F(.)\}$, where Φ is the cumulative standard normal distribution. One major advantage of modeling the multivariate distribution of gestational ages by the normal copula function is that with the monotonic transformation G, the joint distribution can be fully described by arbitrary continuous univariate distributions and a correlation matrix which fully specifies the dependency of a multivariate normal distribution. Normal copula has been used in other settings. For example, Huang and Berry (2006) use normal copula to estimate the association between the mark time and survival time, Song et al. (2009) use normal copula to model the correlation of mixed (continuous and/or discrete) correlated data where the marginal data follow a parametric exponential dispersion family distribution. Othus and Li (2011) assume proportional hazards model for the marginal distribution and use normal copula to model the correlation of multivariate survival data.

In model (4), all the pairs of gestational ages of preterm births share a common correlation. The scatter plot displayed in Figure 1(b), however, suggests that the correlation may vary with the types of preterm birth. That is, the correlation between pairs of gestational ages with PR preterm in both members may be different than that with the other two pair types (PU, PU) and (PR, PU). To accommodate this possibility, model (4) can be extended to model (5),

$$(G(\epsilon_{ik}(y_{ik})), G(\epsilon_{ij}(y_{ij})) \sim \mathbf{\Phi}_{\rho_{y_{ij},y_{ik}}}, \tag{5}$$

where $\rho_{11}, \rho_{22}, \rho_{12}$ correspond to the correlation between gestational ages of PR preterm births in both pregnancies, PU preterm births in both pregnancies, and one type each, respectively.

Models (1)-(5) together determine the joint distribution of pregnancy outcome types and gestational ages of preterm births of all the repeated pregnancies in each woman. They can be used to make inference on a variety of quantities of interest. Two such quantities are 1) the cumulative incidence of PR (PU) preterm birth of the first pregnancy recorded in the study for a set of baseline covariates; 2) the cumulative incidence of PR (PU) preterm birth of a pregnancy given the previous adverse outcome, gestational age and covariates . For brevity of notation, the subscript indexing subject is omitted in the formula displaying these two quantities. The first quantity is formulated as $p(T_1 \leq t, Y_1 = y \mid \mathbf{Z}_1^Y, \mathbf{Z}_1^T)$ which equals $F[t - E(T_1 \mid \mathbf{Z}_1^T, y)]p(Y_1 = y \mid \mathbf{Z}_1^Y)$, where $p(Y_1 = y \mid \mathbf{Z}_1^Y) = \int p(Y_1 = y \mid \mathbf{Z}_1^Y, \mathbf{b})g(\mathbf{b})\, d\mathbf{b}$.

The second quantity for the second pregnancy can be formulated as $p(T_2 \leq t, Y_2 = y_2 \mid a <$

$T_1 \leq b, Y_1 = y_1, \mathbf{Z}_1^T, \mathbf{Z}_1^Y, \mathbf{Z}_2^T, \mathbf{Z}_2^Y), a < b < 37$ weeks, which equals

$$\frac{p(a{<}T_1{\leq}b,T_2{\leq}t|Y_1{=}y_1,Y_2{=}y_2,\mathbf{Z}_1^T,\mathbf{Z}_2^T)p(Y_1{=}y_1,Y_2{=}y_2|\mathbf{Z}_1^Y,\mathbf{Z}_2^Y)}{p(a{<}T_1{\leq}b|Y_1{=}y_1,\mathbf{Z}_1^T)p(Y_1{=}y_1|\mathbf{Z}_1^Y)} =$$

$$\frac{[p(T_1{\leq}b,T_2{\leq}t|Y_1{=}y_1,Y_2{=}y_2,\mathbf{Z}_1^T,\mathbf{Z}_2^T){-}p(T_1{\leq}a,T_2{\leq}t|Y_1{=}y_1,Y_2{=}y_2,\mathbf{Z}_1^T,\mathbf{Z}_2^T)]p_{12}^{y_1,y_2}(\mathbf{Z}_1^Y,\mathbf{Z}_2^Y)}{[p(T_1{\leq}b|Y_1{=}y_1\mathbf{Z}_1^T){-}p(T_1{\leq}a|Y_1{=}y_1,\mathbf{Z}_1^T)]p_1^{y_1}(\mathbf{Z}_1^Y)},$$

where

$$p(T_1 \leq t_1, T_2 \leq t_2 \mid Y_1 = y_1, Y_2 = y_2, \mathbf{Z}_1^T, \mathbf{Z}_2^T) =$$
$$\mathbf{\Phi}[\Phi^{-1}\{F(t_1 - E(T_1 \mid \mathbf{Z}_1^T, y_1))\}, \Phi^{-1}\{F(t_2 - E(T_2 \mid \mathbf{Z}_2^T, y_2))\}],$$

$p(T_1 \leq t \mid Y_1 = y_1, \mathbf{Z}_1^T) = F(t - E(T_1 \mid \mathbf{Z}_1^T, y_1))]$, and $\mathbf{\Phi}(.,.)$ is the bivariate cumulative distribution, and $p_1^{y_1}(\mathbf{Z}_1^Y)$ and $p_{12}^{y_1,y_2}(\mathbf{Z}_1^Y, \mathbf{Z}_2^Y)$ are the probability of $Y_1 = y_1$ and the joint probability of $Y_1 = y_1$ and $Y_2 = y_2$ given the corresponding covariates. Even though T_1 would be known, when it is used in predicting the cumulative incidence of T_2, we use a range for gestational age to allow for the flexibiltiy in relating a range of T_1 to T_2. This is important for our scientific problem. From a population perspective, it is interesting to the obstetrics community to estimate the cumalative incidence of preeclampsia-related preterm birth in a second pregnancy for women who have very early preterm birth in the first pregnancy ($\leq$ 34 weeks). If one is interested in relating a specirfic gestational age of the first pregnancy to that of the second pregnancy (i.e. individual prediction), the value of a can be chosen such that its distance to value b is infinitesimal. In the joint model, outcomes of all the repeated pregnancies in each woman are included, and thus for those with two or more previous pregnancies, we could estimate cumulative incidence of the preterm events using the entire pregnancy history and not just the last pregnancy. For example, for a woman who had two preterm births, it can be useful to predict that woman's risk of PR (PU) preterm birth in her third pregnancy.

3.2 Estimation

In this section, we present an estimation procedure to estimate the parameters in models (1)-(4). We start with the estimation of parameters $\mathbf{\Omega} = (\boldsymbol{\alpha}, \sigma_1, \sigma_2, \nu)$ in the polychotomous logistic model with random effects for the repeated adverse pregnancy outcomes by maximizing its likelihood function given by

$$L(\mathbf{\Omega}) = \prod_i \int \prod_j \prod_{c=0}^{2} p_{ijc}^{I(y_{ij}=c)} g(\mathbf{b})\, db. \tag{6}$$

Numerical integration such as Gauss-Hermite quadrature can be used to perform the integration over the bivariate normal distribution of the correlated random effects specified in (3). Alternatively the EM algorithm may be used to compute the maximum likelihood estimate of $\mathbf{\Omega}$. Based on the standard asymptotic properties of the maximum likelihood estimators, under the correct specification of the random effects polychotomous logistic model, the MLE $\hat{\mathbf{\Omega}}$ is consistent and asymptotically normal with mean equal to the true parameter values $\mathbf{\Omega}$ and variance- matrix Σ which can be consistently estimated by the inverted observed information matrix.

Parameters in models (3)-(4) for gestational age of preterm birth include $\boldsymbol{\gamma} = (\beta_0, \boldsymbol{\beta}, \theta), \rho$, and infinite dimensional F. Joint estimation of $(\boldsymbol{\gamma}, \rho)$ and F is complex. Instead, we extend the work of Klaassen and Wellner (1997) by estimating $\boldsymbol{\gamma}$, ρ and F in different stages . First, we used generalized estimation equation (GEE) under working independence to estimate $\boldsymbol{\gamma}$. Since gestational ages used to fit model (3) are all inside the range of preterm birth (i.e. within 37 weeks), if the model is correctly specified, it is unlikely that the estimated mean gestational age would be outside the range. Hence it is important to check that this condition is satisfied in the estimated model. In the NICHD Consecutive Pregnancies Study data analysis, this scenario did not occur. We then estimated the distribution function F of ϵ by the empirical distribution of the estimated residuals, $\hat{\epsilon}_{ij} = T_{ij} - \hat{\beta}_0 - \hat{\boldsymbol{\beta}}' \mathbf{Z}_{ij}^T - \hat{\theta} I(Y_{ij} = 1)$, given by $\hat{F}(u) = \sum_{i:l_i>0} \sum_j I(\hat{\epsilon}_{ij} \leq u)/N$, where $\hat{\boldsymbol{\gamma}} = (\hat{\beta}_0, \hat{\boldsymbol{\beta}}, \hat{\theta})$ are the GEE estimators, and $N = \sum_i l_i$ is the total number of preterm births. Finally the correlation coefficient ρ was estimated by the normal score rank correlation. Let $\tilde{F} = \frac{N}{N+1} \hat{F}$ denote the rescaled empirical distribution function. The normal scores rank correlation coefficient $\hat{\rho}$ is given by

$$\hat{\rho} = \frac{\frac{1}{N_2 - q - 1} \sum_{i:l_i>1} \sum_{j<k} \Phi^{-1}(\tilde{F}(\hat{\epsilon}_{ij})) \Phi^{-1}(\tilde{F}(\hat{\epsilon}_{ik}))}{\frac{1}{N - q - 1} \sum_{i:l_i>0} \sum_{j=1}^{l_i} [\Phi^{-1}(\tilde{F}(\hat{\epsilon}_{ij}))]^2}, \tag{7}$$

where $N_2 = \sum_{i:l_i>0} l_i(l_i - 1)/2$ and $q = p_2 + 2$.

Based on the GEE asymptotic theory, under the correct specification of the mean function for the gestational age of preterm birth, $\hat{\boldsymbol{\gamma}}$ is consistent for the true parameter values $\boldsymbol{\gamma}$, and as $n_1 \to \infty$, $\sqrt{n_1}(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma})$ converges to zero-normal distribution with sandwich-type variance-covariance, where n_1 is the number of women with at least one preterm birth. The consistency and asymptotic normality of $\hat{F}$ has been established (see Houseman et al. (2004) and references therein). The asymptotic normality of the semi-parametric estimator $\hat{\rho}$ is presented in the following Theorem. The proof follows that of Klaassen and Wellner (1997) for bivariate data without covariates and is sketched in the Appendix.

Theorem 1. Under regularity conditions and given $(\hat{\boldsymbol{\gamma}}, \tilde{F})$ are consistent and asymptotically normal for $(\boldsymbol{\gamma}, F)$ and the normal copula model (4) holds for ϵ_{ij}, the estimator $\hat{\rho}$ of ρ is consistent, and as $n_2 \to \infty, \sqrt{n_2}(\hat{\rho} - \rho)$ converges weakly to a zero-mean normal distribution with variance τ^2, where n_2 is the number of subjects with more than one preterm birth.

Derivation for τ^2 is supplied in the Appendix. The variance τ^2 is a function of $(\boldsymbol{\beta}, \theta, F)$ and the density f of F. An estimator of τ^2 can be obtained by inserting $(\hat{\boldsymbol{\beta}}, \hat{\theta}, \tilde{F})$ for $(\boldsymbol{\beta}, \theta, F)$ and a non-parametric estimate, e.g. kernel density estimate, for f.

The above estimator of ρ can be adjusted to estimate the three types of correlation coefficient in model (5). Specifically,

$$\hat{\rho}_{lm} = \frac{\frac{1}{N_{lm}-q-1}\sum_{i:l_i>1}\sum_{(y_{ij},y_{ik})=(l,m)\cup(m,l)} \Phi^{-1}(\tilde{F}(\hat{\epsilon}_{ij}))\Phi^{-1}(\tilde{F}(\hat{\epsilon}_{ik}))}{\frac{1}{N-q-1}\sum_{i:l_i>0}\sum_{j=1}^{l_i}[\Phi^{-1}(\tilde{F}(\hat{\epsilon}_{ij}))]^2}, \tag{8}$$

where $N_{lm} = \sum_{i:l_i>1}\sum_{(y_{ij},y_{ik})=(l,m)\vee(m,l)}$ is the number of pairs with the adverse pregnancy outcomes equal to (l, m) or (m, l). The asymptotic property of the estimator (8) follows Theorem 1.

4. ANALYSIS OF NICHD CONSECUTIVE PREGNANCIES STUDY DATA

We began by fitting the random effects polychotomous logistic model, where the covariates included history of preterm birth, parity (> 0 versus $=0$), number of fetuses (> 1 versus $=1$), body mass index (BMI) in the beginning of each pregnancy, maternal age, chronic hypertension (yes verse no), smoking (yes versus no). Of these covariates, chronic hypertension and smoking were subject-specific collected at the beginning of the study, and the rest of the covariates were pregnancy-specific collected at each pregnancy. We treated smoking as a subject-specific covariate assessed at entry into the cohort, because we were interested in long-term smoking status rather than short-term for women who suddenly stopped smoking, possibly due to an abnormal prior pregnancy. In addition to these covariates, some interaction terms were also considered in the model. Particularly, a woman with a history of preterm birth must have positive parity, and hence the effect of history of preterm birth must be estimated through the interaction term of history of preterm birth and parity status (>0 versus 0). In addition, the covariate effects on the pregnancy outcomes of single fetus pregnancies (singleton) may be different than those of multiple fetus pregnancies (twins, triplets, etc). It is not feasible to do a stratified analysis on singleton versus multiple births, since each woman can potentially have both singleton and multiple births in her repeated pregnancies. To deal with this issue, we added the interaction terms between all other covariates and number of fetuses (> 1) to the model. The estimates were obtained by maximizing (2) where the double integrals were approximated by the product of 10-point one-dimensional Gauss-Hermite quadrature. With the exception of the interactions between a history of preterm birth and parity and between BMI and number of fetuses, no other interaction terms were significant, and the parameter estimates in the final model are listed in Table 3. For PR preterm birth, with the exception of maternal age, all the other covariates were significant ($p<0.05$). Among them, number of fetuses (> 1 versus $=1$) had the largest effect on the probability of having PR preterm birth. For a woman with BMI=24 kg/m^2, the odds of a PR preterm birth compared to a term birth was 51 times higher ($e^{-.086\times 24+5.996}$) with two or more fetuses than with single fetus. Increasing BMI, chronic hypertension, being a smoker, and history of preterm birth all increased the likelihood of having a PR preterm birth, whereas parity was negatively associated with PR preterm birth: the odds of a PR preterm birth for parous women (parity > 0) was 25% ($e^{-1.395}$) of that for nulliparous women (parity $=0$). For PU preterm birth, all the covariates listed in Table 3 were significant except the interaction between number of fetuses and BMI. Number of fetuses was also the strongest predictor. Although chronic hypertension was also a significant predictor of PU preterm birth, the odds-ratio was 56% ($e^{0.79}/e^{1.37} = 0.56$) of that for PR preterm birth, Parity was also negatively associated with PU preterm birth, but its effect on PU preterm birth was much smaller than on PR preterm birth. The standard deviations of the random intercepts for both PR and PU preterm birth were large, implying high correlation of developing the same type of preterm birth in repeated pregnancies. In contrast, the correlation of the two random intercepts was negative and nearly zero, implying that the correlation of these two types of preterm birth was negligible.

In the next step, we estimated model (3) and (5) for repeated gestational ages of preterm births. Model (3) was estimated by GEE under a working independence assumption. Maternal age and BMI were not significant and were excluded from model (3). The interaction between history of preterm birth and number of fetuses was negatively correlated with gestational age. The estimates are listed in Table 4. If a woman had a history of preterm birth and had more than 1 fetus, the mean gestational age was shortened by almost 2 weeks compared to not having these conditions. The gestational age for PU birth on average was 0.35 week longer than PR preterm birth. The normal scores rank correlation equaled $(0.47, 0.18, 0.07)$ for gestational ages of PR preterm birth in both pregnancies, PU preterm birth in both pregnancies, and one each type, respectively confirming our initial observation (Figure 1(b)). This indicates that the correlation between gestational ages of PR preterm birth is much stronger than either of PU preterm birth or of different types of preterm birth.

[Table 3 about here.]

[Table 4 about here.]

Individualized estimation of the cumulative incidence function is one of primary interests. The cumulative incidence of PR preterm birth with respect to parity and history of preterm birth is plotted in Figure 3(a), where the other covariates used to compute the cumulative incidence were set at their median values (maternal age=27.1 years, BMI=24.8 kg/m^2, number of fetuses=1, chronic hypertension=no, smoker=no). The figure shows that the cumulative incidence is highest for nulliparous women and lowest for parous women with no history of preterm birth. The risk of PR preterm birth for women with a history of preterm birth was almost identical to that for nulliparous women. The cumulative incidence of PU preterm birth for the same set of covariate values is plotted in Figure 3(b). It is highest for women with a history of preterm birth and lowest for parous women with no history of preterm birth.

[Figure 3 about here.]

The cumulative incidence of any recurrent preterm (PR and PU) birth for parous women with history of preterm birth at the previous pregnancy is plotted in Figures 4, where the solid line is the marginal cumulative incidence of PR (PU) preterm birth for women with history of preterm birth. In these figures, the dashed and dotted line correspond to the conditional cumulative incidence given the PR (or PU) preterm birth before 32 weeks and between 32 and < 37 weeks of gestational age at the first pregnancy, respectively. The covariate values used to compute the conditional cumulative incidence were set the same as before except maternal age = 29 years. A few patterns are commonly observed in the four figures. First, the conditional cumulative incidence is higher than the marginal cumulative incidence if the type of recurrent preterm birth is the same as that in the previous pregnancy. If the preterm birth of the two pregnancies are different types (Figure 4(b)-4(c)), the conditional cumulative incidence is slightly lower than the marginal counterpart. This is due to the negative correlation coefficient estimate of the random intercepts. However, because the correlation coefficient estimate is not significant, the 95% confidence intervals of these cumulative incidences overlap. Second, the risk of recurrent PR (PU) preterm birth is higher if the gestational age of previous preterm birth is less than 32 weeks than if the gestational age is between 32 and < 37 weeks. The 95% pointwise confidence intervals plotted in Figures 3 and 4 were obtained from the 2.5 and 97.5 percentiles of 250 bootstrap samples, where the sampling unit is woman participant.

[Figure 4 about here.]

We checked the modelling assumptions by comparing the observed types and timing of preterm birth versus their predicted counter parts. In model (1), after dichotomizing the continuous covariates age and BMI at their respective medians, the data set was grouped according to the unique combination of the 8 binary covariates. In each group, the mean observed proportion of PR (PU) preterm birth was calculated. Since the incidence of preterm birth is low, the observed mean proportions for groups with less than 200 observations were highly variable and excluded from the comparison. The observed mean versus predicted proportion of PR and PU preterm birth are displayed in Figure 5(a) and 5(b), respectively. Overall, there is high agreement between the observed and predicted proportions, indicating model (1) fits the data well. We also compared the number of women having (PR,PR), (PR,PU),(PU,PR), and (PU,PU) in their first two pregnancies against the predicted counterparts. The discrepancy is within 10% indicating model (1) describes the dependency structure in outcome types across pregnancies well. For model (5), all the covariates are binary. We repeated the same procedure described above and plotted the observed mean gestational age versus the predicted gestational age of preterm birth, where groups with less than 15 observations were excluded. The scatter plot fluctuating around the 45 degree line indicates that model (5) fits the distribution of gestational age of preterm birth well. Since the correlation between gestational ages of repeated preterm births was modeled through a semi-parametric normal copula model, and the transformed normal rank scores by definition are multivariate normal, model checking is not necessary for the correlation.

[Figure 5 about here.]

5. SIMULATION STUDY

We conducted a simulation study to evaluate the performance of the estimation procedure presented in Section 3. The parameter settings in the simulation study were chosen to mimic the NICHD study characteristics, including the number of pregnancies, the low baseline rates of PR and PU preterm birth, and higher correlation between repeated PR preterm births than repeated PU preterm births. We first generated the correlated event types $\mathbf{Y}_i = (Y_{i1}, \cdots, Y_{ik_i})$ from the polychotomous logistic regression with random effects model (1), where the number of repeated pregnancies k_i was random ranging from 2 to 6 with probability (0.7, 0.15,0. 1,0.045, 0.005), the fixed-effect intercepts $(\alpha_{01}, \alpha_{02}) = (-3.81, -2.42)$ corresponding to baseline event rate of 0.02 and 0.08 for PR and PU preterm birth, respectively, slopes for one fixed pregnancy-specific binary covariate X_{ij}, $(\alpha_1, \alpha_2) = (-1, -0.5)$, and the random intercepts (b_{i1}, b_{i2}) follows a bivariate zero-mean normal distribution with standard deviations $\sigma_1 = 2$, $\sigma_2 = 1.5$, and correlation coefficient $\nu = 0.5$. Gestational age was assumed to follow a normal distribution with mean equal to $\mu_{ij} = \beta_0 + \beta_1 X_{ij} + \theta I(Y_{ij} = 2)$ and standard deviation SD=4. Correspondingly, the gestational age of preterm birth was conditional normal given $T_{ij} \leq 37$ and the correlation of the gestational ages of preterm birth from the same woman was specified through the multivariate normal copula function. Specifically, $T_{ij} \mid T_{ij} \leq 37$, or equivalently $T_{ij} \mid Y_{ij} > 0$, was generated by $\Phi^{-1}(u_{ij}\Phi\{(37 - \mu_{ij})/SD\}) \times SD + \mu_{ij}$. where u_{ij}'s are multivariate standard normal with common correlation coefficient $\rho = 0.3$. Since the mean of the conditional normal distribution of the gestational age of preterm birth was no longer equal to μ_{ij}, a saturated linear model was fitted with the mean function $\mu^*_{ij} = \beta^*_0 + \beta^*_1 X_{ij} + \theta^* I(Y_{ij} = 2) + \gamma X_{ij} \times I(Y_{ij} = 2)$. The sample size was set at 20,000 and the simulations were repeated 500 times

Due to the low preterm birth rate, one of the standard deviations of the random intercepts converged to the boundary value of zero in 7% of the simulations. This problem is alleviated when the preterm birth rate is set at a higher value (data not shown). The simulation results with positive standard deviation of the random intercept are summarized in Table 5. As the Monte-Carlo means of all the parameter estimates were close to the true values, there was little bias in the estimation. The Monte-Carlo standard errors are close to the square root of the Monte-Carlo means of variances which were obtained by inverting the numerically approximated hessian matrix, and the coverage probabilities for most of the parameters were close to the 95% nominal level. The correlation coefficient ρ in the normal copula model was estimated by the normal score rank correlation coefficient given in (7). If the correlation coefficient were estimated using the untransformed residuals $\hat{\epsilon}_{ij}$s, the estimator would have negative bias with the Monte-Carlo mean equal to 0.28.

[Table 5 about here.]

6. DISCUSSION

We proposed a joint model for individualized risk prediction of adverse outcomes in repeated pregnancies when the outcomes are subject to other competing events. In this article, we develop a novel statistical methodology and apply it to a unique data set in order to solve an important problem in obstetrics. The estimates appear to be obtained by a two-step procedure, but the estimates so obtained for model (1) would be identical to the estimates obtained from the joint likelihood. This is because the parameters in model (1) do not appear in model (3),and the likelihood specified in (6) is proportional to the likelihood of the joint model. The estimation of parameters in model (3) involves the infinite-dimensional distribution function of residual gestational age, and the likelihood-based approach would be computationally challenging. For this reason, we proposed a plug-in based approach for the estimation of model (3) parameters. Since model (3) does not involve parameters in model (1), as long as the model is correctly specified, according to *Theorem 1*, the parameter estimates are consistent.

In the analysis, we focus on the examination of risk factors for PR preterm birth, where preterm birth for other reasons is a competing event. The results will have important implications for managing women with a past history of preeclampsia. There are many other important adverse outcomes that are competing risks, such as preterm birth due to spontaneous labor where medically indicated preterm births would be a competing event. We plan to use this methodology to estimate incidence curves for this type of preterm birth in a subsequent publication in the medical literature.

There is a limited literature on competing risks for correlated time to event data. Bandeen-Roche and Liang (2002) considered a frailty model for correlated failure times subject to competing risks, and introduced a non-parametric cause-specific hazard ratio association measure for bivariate competing risk data. Shih and Albert (2010) proposed a bivariate model for correlated event-times subject to competing risk by incorporating association between times to first events and associations between failure types given the first event times. However, extension of their proposed model to multivariate competing risk survival data is complex and yet to be developed. Gorfine and Hsu (2011) proposed a frailty-based proportional hazards competing risks model for multivariate survival data, where cause-specific frailty processes are used to induce the association between cause-specific failure times. However, these approaches do not directly address many scientific questions in the NICHD Consecutive Pregnancies Study. In the proposed approach, we model the dual outcome of the occurrence and type of preterm birth as well as the gestational ages for those pregnancies that are preterm. Specifically, the proposed approach directly (i) estimates regression relationships between important subject and pregnancy-specific covariates and the risk of different types of preterm birth (PR and PU preterm birth) and (ii) estimates the effect of these covariates on gestational age for births that are preterm. Further, the proposed approach allows us to estimate and directly interpret the correlation in these two components (occurrence of preterm birth and timing of preterm birth) across consecutive pregnancies. Furthermore, similar to both Shih and Albert (2010) and Gorfine and Hsu (2011), we are able to estimate cumulative incidence functions, which is also an important objective.

In both model (1) and model (3), we assume exchangeable correlation between repeated pregnancy outcomes (e.g., the occurrence of preterm birth type and the timing of gestational age for a preterm birth pregnancy each have an exchangeable correlation structure across repeated pregnancies.). This means that the correlation in the outcomes from different pregnancies does not depend on the distance between these pregnancies (e.g., 1st and 2nd or 1st and 3rd). To evaluate whether consecutive and non-consecutive preterm births have similar correlation, we collapsed the two types of pre-term births and examined the frequencies of pre-term births in the first 3 pregnancies. Ninety-eight women had preterm birth, followed by a term birth, followed by another preterm birth, and ninety-nine women had preterm birth, followed by another pre-term birth, followed by a term birth. The fact that the two preterm birth patterns had an almost identical frequency indicates that the exchangeable correlation structure for model (1) is reasonable. The correlation of the gestational ages of the 98 non-consecutive preterm births and 99 consecutive pre-term births are 0.173 and 0.167, respectively. The similarity in the two correlations indicates that the exchangeability assumption for model (3) is reasonable as well. However, for some adverse birth outcomes, a serial correlation may be more appropriate; in this case, the correlation may be stronger for pregnancies that occurred closer together either in time or in order. Serial correlation can be incorporated in a number of ways including the introduction of a shared random process rather than a shared random effect. We would need to employ Monte-Carlo EM or other numerically intensive methods for estimation in this case.

The modeling framework also assumes that the number of pregnancies is not related to the occurrence of adverse outcomes or gestational age at the repeated births. We examined this assumption with some simple data analysis. For example, a simple plot of the proportion of pre-term births versus number of pregnancies showed no overall pattern (data not shown). In addition, a similar plot of the average patient-specific gestational age versus number of pregnancies also showed no pattern. Accounting for an informative number of pregnancies (those with an adverse outcome are more likely to have a larger number of pregnancies over this fixed nine year interval) is an area of future research.

APPENDIX

Derivation of the asymptotic distribution of $\hat{\rho}$.

Since $(\hat{\gamma}, \tilde{F})$ are consistent for (γ, F), the denominator of $\hat{\rho}$ is asymptotically equivalent to $\frac{1}{N-q-1}\sum_{i:l_i>0}\sum_{j=1}^{l_i}[\Phi^{-1}(F(\epsilon_{ij}))]^2$ which converges to one as $\Phi^{-1}(F(\epsilon_{ij})$ is a standard normal random variable. Hence $\sqrt{n_2}(\hat{\rho}-\rho)$ is asymptotically equivalent to

$\sqrt{n_2}\{\frac{1}{N_2}\sum_{i:l_i>1}\sum_{j<k}\Phi^{-1}(\tilde{F}(\hat{\epsilon}_{ij}))\Phi^{-1}(\tilde{F}(\hat{\epsilon}_{ik}))-\rho\}$ which can be rewritten as

$$\begin{aligned}
\sqrt{n_2}(\hat{\rho}-\rho) &\approx \frac{\sqrt{n_2}}{N_2}\sum_{i:l_i>1}\sum_{j<k}\Phi^{-1}(\tilde{F}(\hat{\epsilon}_{ij}))\Phi^{-1}(\tilde{F}(\hat{\epsilon}_{ik}))-\sqrt{n_2}\rho \\
&= \frac{\sqrt{n_2}}{N_2}\sum_{i:l_i>1}\sum_{j<k}[\Phi^{-1}(\tilde{F}(\hat{\epsilon}_{ij}))-\Phi^{-1}(F(\epsilon_{ij}))]\Phi^{-1}(F(\epsilon_{ik}))+ \\
&\frac{\sqrt{n_2}}{N_2}\sum_{i:l_i>1}\sum_{j<k}[\Phi^{-1}(\tilde{F}(\hat{\epsilon}_{ik}))-\Phi^{-1}(F(\epsilon_{ik}))]\Phi^{-1}(F(\epsilon_{ij}))+ \\
&\frac{\sqrt{n_2}}{N_2}\sum_{i:l_i>1}\sum_{j<k}\left\{\Phi^{-1}(F(\epsilon_{ij}))\Phi^{-1}(F(\epsilon_{ik}))-\rho\right\}+o_p(1). \qquad \text{(A.1)}
\end{aligned}$$

The first term in the above equation by Taylor's expansion and integration by parts can be rewritten as

$$\begin{aligned}
&\frac{\sqrt{n_2}}{N_2}\sum_{i:l_i>1}\sum_{j<k}[\Phi^{-1}(\tilde{F}(\hat{\epsilon}_{ij}))-\Phi^{-1}(F(\epsilon_{ij}))]\Phi^{-1}(F(\epsilon_{ik})) \\
&= \frac{\sqrt{n_2}}{N_2}\sum_{i}\sum_{j<k}\left\{\frac{1}{\phi(\Phi^{-1}(F(\epsilon_{ij})))}(\tilde{F}-F)(\epsilon_{ij})-\frac{f(\epsilon_{ij})}{\phi(\Phi^{-1}(F(\epsilon_{ij})))}\mathbf{X}_{ij}^t(\hat{\boldsymbol{\gamma}}-\boldsymbol{\gamma})\right\}\times \\
&\quad \Phi^{-1}(F(\epsilon_{ik}))+o_p(1) \\
&= \sqrt{n_2}\int\left\{\frac{1}{\phi(\Phi^{-1}(F(u)))}(\tilde{F}-F)(u)-\frac{f(u)}{\phi(\Phi^{-1}(F(u)))}\mathbf{X}^t(\hat{\boldsymbol{\gamma}}-\boldsymbol{\gamma})\right\}\times \\
&\quad \Phi^{-1}(F(v))\,dP_n((u,\mathbf{X}),v)+o_p(1) \\
&= \sqrt{n_2}\Bigg\{\int\frac{\Phi^{-1}(F(v))}{\phi(\Phi^{-1}(F(u)))}(\tilde{F}-F)(u)dP_{\rho,F}(u,v)-\int\frac{\mathbf{X}^t f(u)\Phi^{-1}(F(v))}{\phi(\Phi^{-1}(F(u)))}\,dP((u,\mathbf{X}),v)\times \\
&\quad (\hat{\boldsymbol{\gamma}}-\boldsymbol{\gamma})\Bigg\}+o_p(1) \\
&= \sqrt{n_2}\left\{\int(\tilde{F}-F)(u)\rho\Phi^{-1}(F(u))\,d\Phi^{-1}(F(u))-\boldsymbol{\xi}^t(\hat{\boldsymbol{\gamma}}-\boldsymbol{\gamma})\right\}+o_p(1) \\
&= \frac{-\rho}{2\sqrt{n_2}}\left\{\frac{n_2}{N}\sum_{i}\sum_{j}(\Phi^{-1}(F(\epsilon_{ij}))^2-1)\right\}-\sqrt{n_2}\boldsymbol{\xi}^t(\sum_{i:l_i>0}\mathbf{X}_i^t\mathbf{X}_i)^{-1}\{\sum_{i}\mathbf{X}_i^t(\mathbf{Y}_i-\mathbf{X}_i\boldsymbol{\gamma})\}+ \\
&\quad o_p(1) \\
&\quad \frac{1}{\sqrt{n_2}}\sum_{i:l_i>0}\psi_i+o_p(1),
\end{aligned}$$

where $\mathbf{X}_{ij}=\{1,\mathbf{Z}_{ij}^T,I(Y_{ij}=1)\}$,$\mathbf{X}_i=(\mathbf{X}_{i1}^t,\cdots,\mathbf{X}_{il_i}^t)^t$, $P_{\rho,F}$ is the bivariate normal copula model with the common marginal distribution F and correlation coefficient ρ, $P_n((u,\mathbf{X}),v)$ is the empirical c.d.f. with mass points $((\epsilon_{ij},\mathbf{X}_{ij}),\epsilon_{ik})$ of equal mass $1/N_2$, P is the converging distribution function of $P_n((u,\mathbf{X}),v)$, $\boldsymbol{\xi}=\int\frac{\mathbf{X}f(u)}{\phi(\Phi^{-1}(F(u)))}\Phi^{-1}(F(v))\,dP((u,\mathbf{X}),v)$, and

$$\psi_i=\frac{-\rho n_2}{2N}\sum_{j}(\Phi^{-1}(F(\epsilon_{ij}))^2-1)-\boldsymbol{\xi}^t(n_2^{-1}\sum_{i:l_i>0}\mathbf{X}_i^t\mathbf{X}_i)^{-1}\{\mathbf{X}_i^t(\mathbf{Y}_i-\mathbf{X}_i\boldsymbol{\gamma})\}$$

.

The ψ_i's are n_2 i.i.d. random variables, and thus the first term in (A.1) converges to a normal distribution with mean 0 and variance $E(\psi_1^2)$. The second term in (A.1) has the same asymptotic distribution as the first term, and the third term is asymptotically equivalent to $n_2^{-1/2}\sum_{i:l_i>0}\varphi_i$ which is the sum of n_2 i.i.d. random variables with $\varphi_i = (n_2/N_2)\sum_{j<k}\Phi^{-1}(F(\epsilon_{ij}))\Phi^{-1}(F(\epsilon_{ik})) - \rho$. It follows that $\sqrt{n_2}(\hat{\rho} - \rho)$ converges to to normal distribution with mean 0 and variance equal to $\tau^2 = E[(2\psi_1 + \varphi_1)^2]$.

REFERENCES

Ananth, C. V., Getahun, D., Peltier, M. R., Salihu, H. M., and Vintzileos, A. M. (2006). Recurrence of spontaneous versus medically indicated preterm birth. *American Journal of Obstetrics and Gynecology*, 195:643–650.

Bandeen-Roche, K. and Liang, K. (2002). Modelling multivariate failure time associations in the presence of a competing risk. *Biometrika*, 89:299–314.

Gorfine, M. and Hsu, L. (2011). Frailty-based competing risks model for multivariate survival data. *Biometrics*, 67:415–426.

Houseman, E. A., Ryan, L. M., and Coull, B. A. (2004). Cholesky residuals for assessing normal errors in a linear model with correlated outcomes. *Journal of the American Statistical Association*, 99:383–394.

Huang, Y. and Berry, K. (2006). Semiparametric estimation of marginal mark distribution. *Biometrika*, 93:895–910.

Kalbfleisch, J. and Prentice, R. (2002). *The Statistical Analysis of Faiture Time Data*. New York: Wiley, 2nd edition edition.

Klaassen, C. A. and Wellner, J. A. (1997). Efficient estimation in the bivaraite normal copula model: normal margins are least favourable. *Bernoulli*, 3:55–77.

Laughon, K., Albert, P., Leishear, K., and Mendola, P. (2013). The nichd consecutive pregnancies study: Recurrent preterm delivery by subtype. *American Journal of Obstetrics & Gynecology*. To appear.

Louis, G. M., Dukic, V. M., Heagerty, P. J., Louis, T. A., Lynch, C. D., Ryan, L. M., Schisterman,

E. F., A, T., and the Pregnancy Modeling Working Group (2006). Statistical issues in modeling pregnancy outcome data. *Statistical Methods in Medical Research*, 15:3–126.

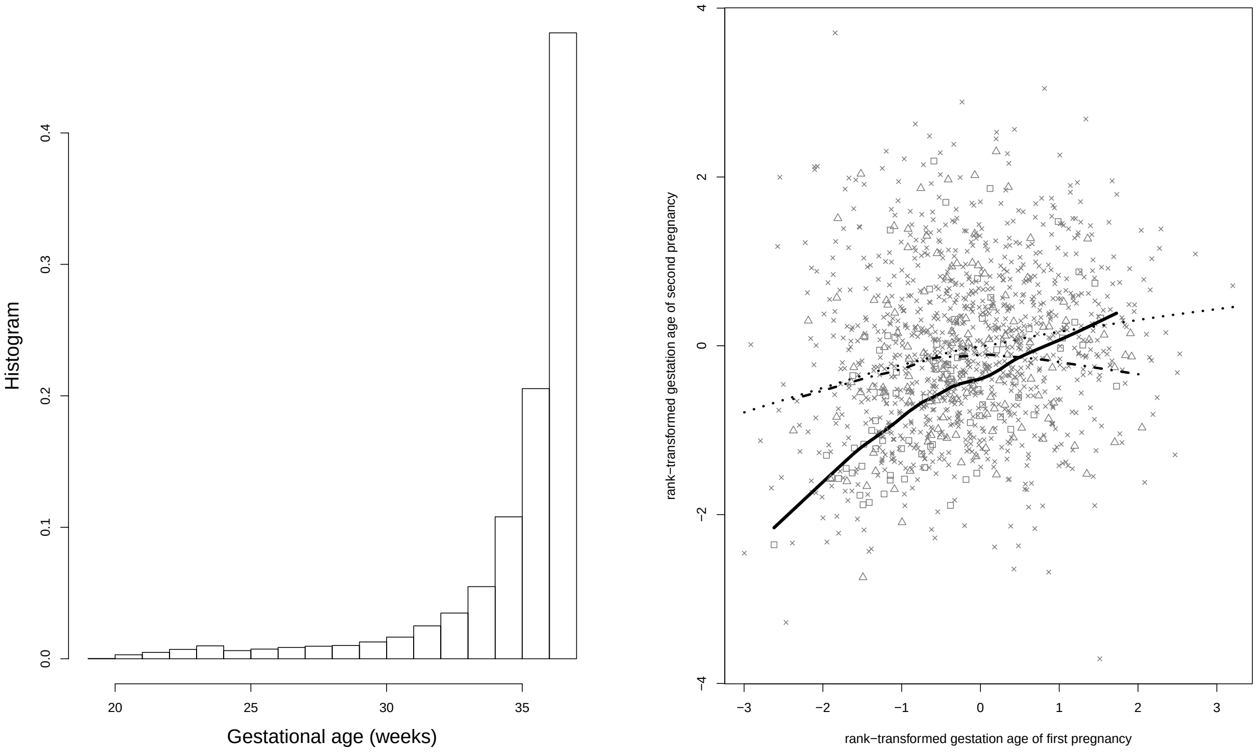
Othus, M. and Li, Y. (2011). A gaussian copula model for multivariate survival data. *Statistics in Biosciences*, 2:154–179.

Shih, J. and Albert, P. (2010). Modeling familiar association of ages at onset of disease in the presence of competing risk. *Biometrics*, 66:1012–1023.

Song, P. X.-K., Li, M., and Ying, Y. (2009). Joint regression analysis of correlated data using gaussian copulas. *Biometrics*, 65:60–68.

Whittemore, A. (1995). Logistic regression of family data from case-control studies. *Biometrika*, 82:57–67.

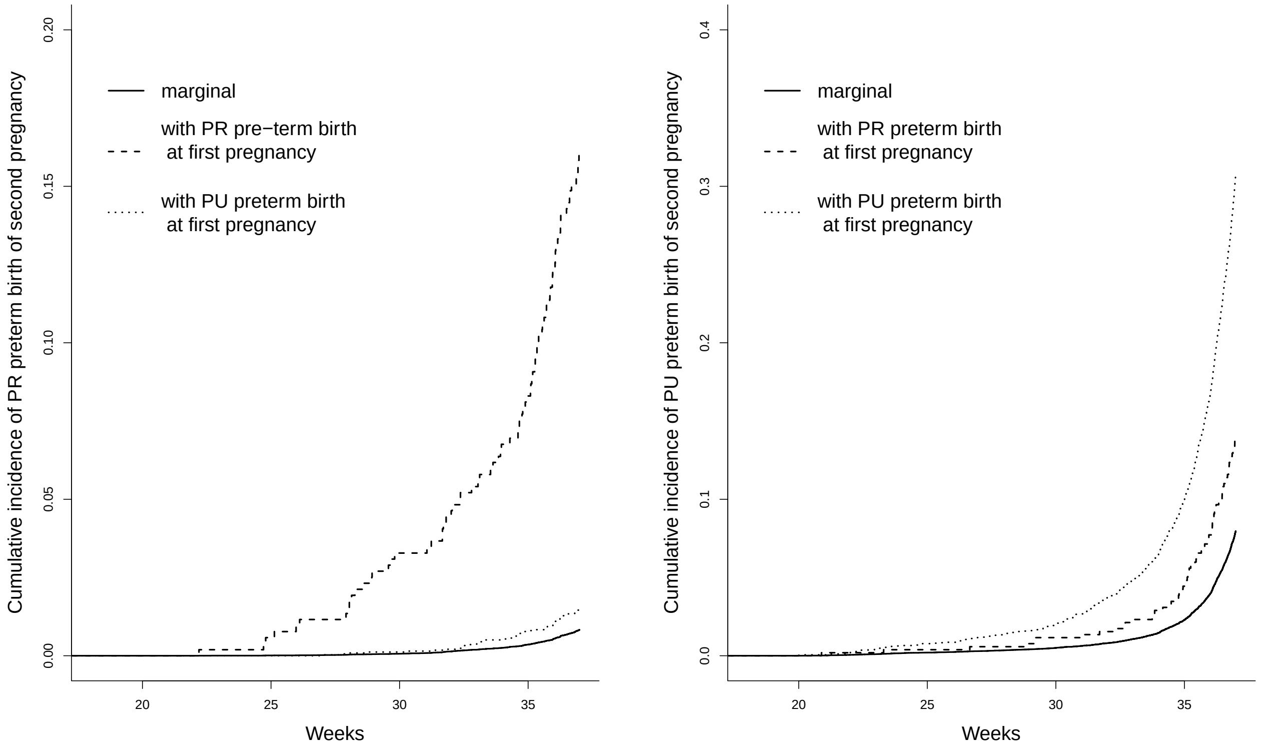
List of Figures



(a)

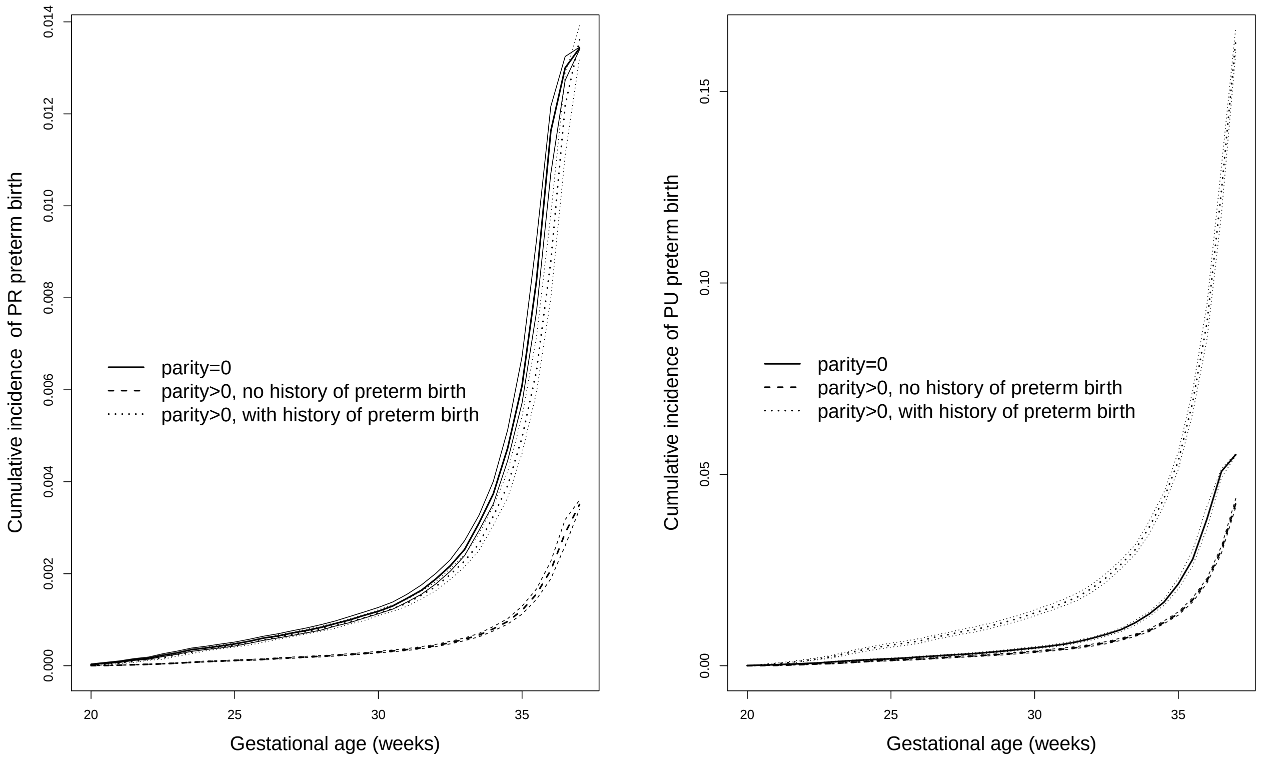
(b)

Figure 1: Histogram and scatter plot of gestational age: (a) Histogram of gestational age of preterm birth; (b) Scatter plot of the normal rank score of gestational age of the first pregnancy versus normal rank score of gestational age of the second pregnancy among women who delivered preterm births in both pregnancies, where the squares and solid lowess smooth correspond to PR preterm births in both pregnancies, cross and dotted lowess smooth to PU preterm births in both pregnancies, and triangles and dotdashed lowess smooth to one of each type.



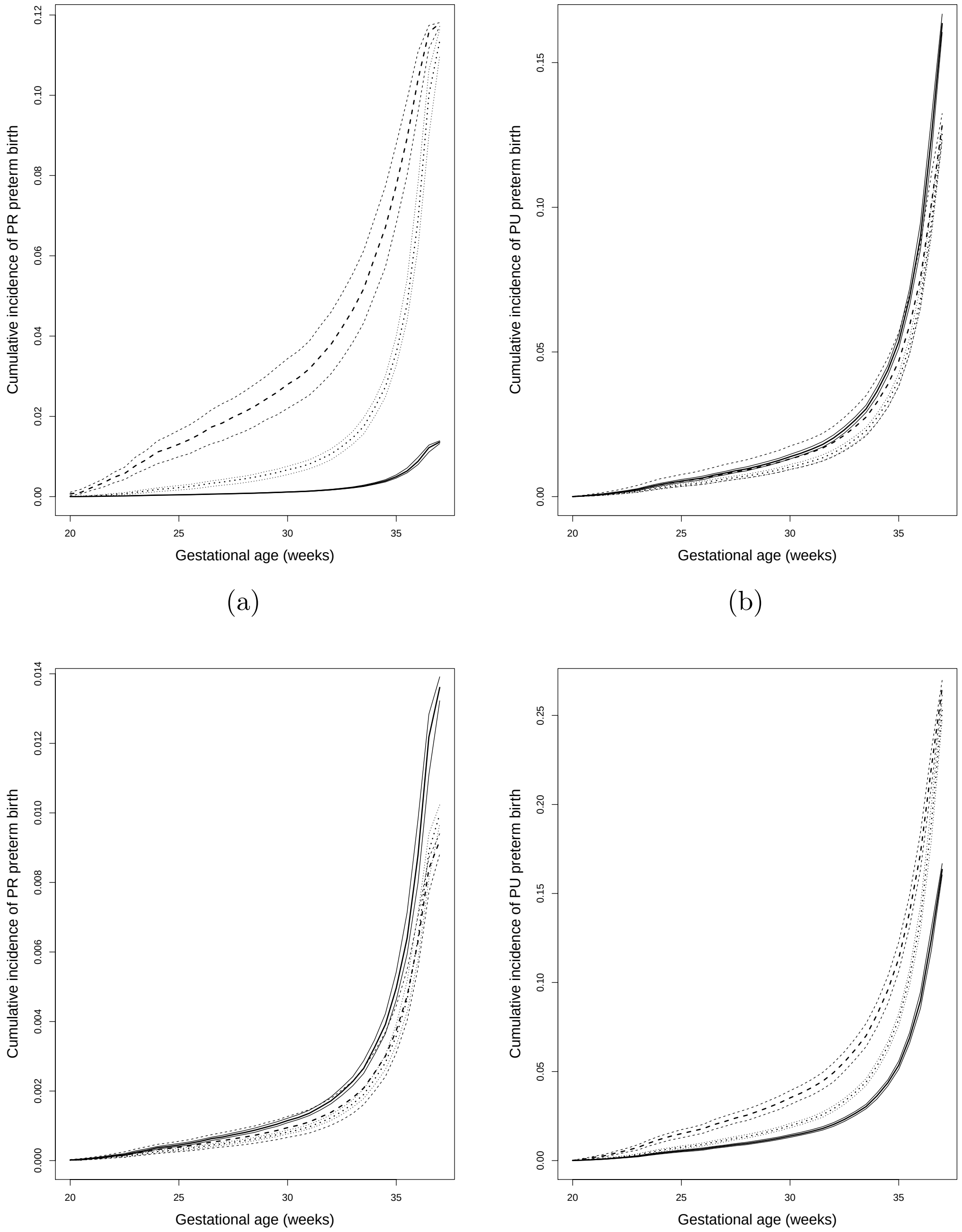
(a) (b)

Figure 2: Cumulative incidence curve of the second preterm birth: (a) The solid, dashed and dotted line corresponds to marginal cumulative incidence of PR preterm birth, cumulative incidence of PR preterm birth given PR pre-term birth at first pregnancy and cumulative incidence of PR preterm birth given PU pre-term birth at first pregnancy, respectively. (b) The solid, dashed and dotted line corresponds to marginal cumulative incidence of PU preterm birth, cumulative incidence of PU preterm birth given PR pre-term birth at first pregnancy and cumulative incidence of PU preterm birth given PU pre-term birth at first pregnancy, respectively.



(a) (b)

Figure 3: Individualized cumulative incidence curve of PR and PU preterm birth: (a) The solid, dashed and dotted line correspond to cumulative incidence and its 95% pointwise confidence interval of PR preterm birth for nulliparous woman, parous woman with no history of preterm birth and parous women with history of preterm birth, respectively; (b) The solid, dashed and dotted line correspond to cumulative incidence and its 95% pointwise confidence interval of PU preterm birth for nulliparous woman, parous woman with no history of preterm birth and parous women with history of preterm birth, respectively.



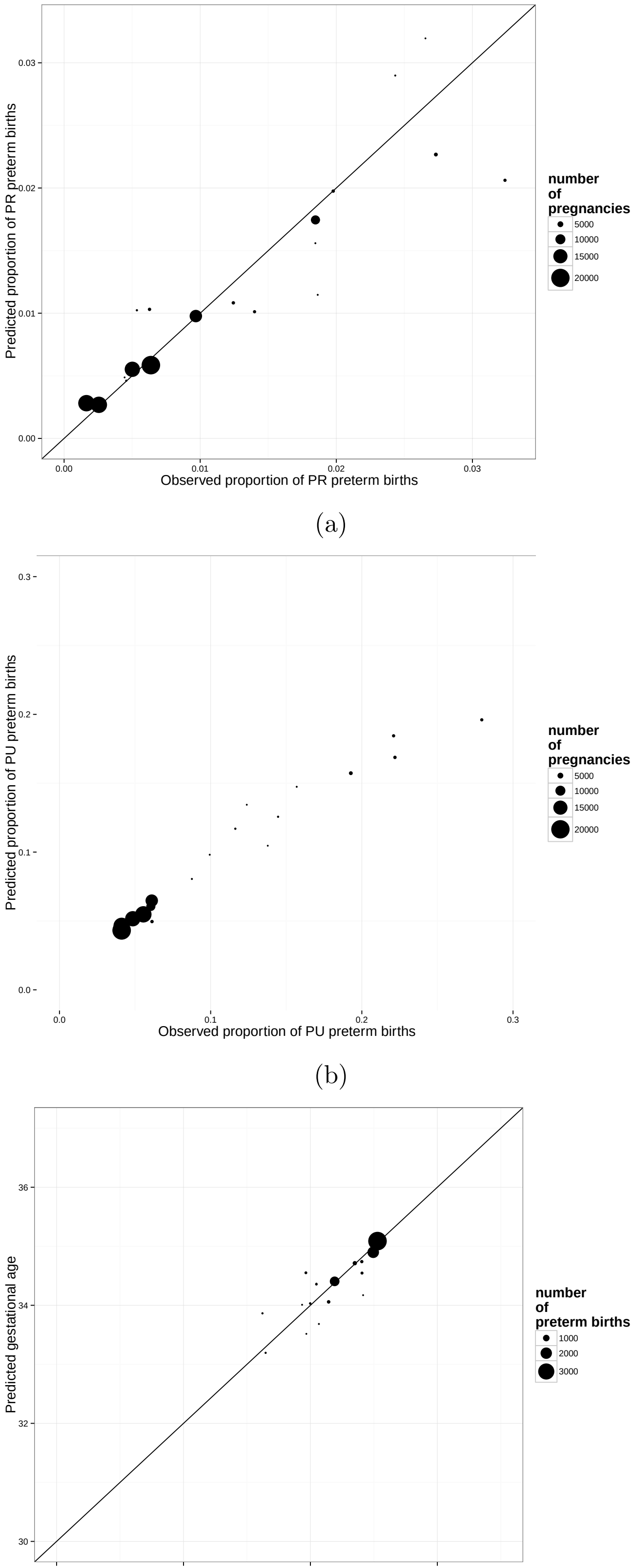
(a)

(b)

(c)

(d)

Figure 4: Individualized cumulative incidence curve and 95% confidence interval of any recurrent preterm birth: (a)-(b) The solid line is the marginal cumulative incidence, and the dashed and dotted line correspond to the conditional cumulative incidence given the PR preterm birth before 32 weeks and between 32 and $<$ 37 weeks of gestational age at the previous pregnancy, respectively; (c)-(d) the solid line is the marginal cumulative incidence, and the dashed and dotted line correspond to the conditional cumulative incidence given the PU preterm birth before 32 weeks and between 32 and $<$ 37 weeks of gestational age at the previous pregnancy, respectively



(a)

(b)

(c)

Figure 5: Mean observed versus predicted type and timing of preterm birth. (a) Mean observed versus predicted proportion of PR preterm birth; (b) Mean observed versus predicted proportion of PU preterm birth; (c) Mean observed versus predicted gestational age of preterm birth.

List of Tables

Table 1: Tabulation of woman participants by the number of pregnancies and preterm births

Number of pregnancies	Number of preterm births						Total
	0	1	2	3	4	5	
2	34255	4711	995	-	-	-	39961
3	8012	1313	337	125	-	-	9787
4	950	190	60	34	11	-	1245
5	53	9	6	1	1	1	71
6	2	0	0	0	0	0	2
Total	43272	6223	1398	160	12	1	51066

Table 2: Tabulation of women with multiple adverse pregnancy outcomes by preeclampsia-related and preeclampsia-unrelated birth

Number of	Number of preeclampsia-unrelated preterm birth					
preeclampsia-related	0	1	2	3	4	Total
0	-	-	1165	140	10	1315
1	-	143	12	0	0	155
2	90	2	1	1	0	94
3	6	1	0	0	0	7
Total	96	146	1178	141	10	1571

Table 3: Parameter estimates of model (1)

	PR related preterm birth			PU preterm birth		
	Estimate	SE	95% CI	Estimate	SE	95% CI
intercept	-7.871	0.311	(-8.481, -7.26)	-2.205	0.098	(-2.397,-2.014)
mothers' age	0.004	0.009	(-0.013, 0.021)	-0.027	0.003	(-0.033,-0.021)
BMI	0.083	0.006	(0.072, 0.095)	-0.008	0.002	(-0.013, -0.003)
parity (> 0)	-1.395	0.088	(-1.568, -1.222)	-0.139	0.034	(-0.205,-0.073)
# of fetus (> 1)	5.996	0.682	(4.658, 7.333)	3.813	0.317	(3.192,4.434)
chronic hypertension (Yes)	1.37	0.194	(0.989, 1.75)	0.79	0.107	(0.58,0.999)
smoker (Yes)	0.664	0.185	(0.303, 1.026)	0.928	0.061	(0.809, 1.048)
parity× history of preterm birth	1.578	0.114	(1.354, 1.802)	1.555	0.042	(1.474,1.637)
# of fetus × BMI	-0.086	0.026	(-0.137, -0.036)	-0.004	0.012	(-0.028, 0.02)
	Estimate	SE	95% CI			
σ_1	1.801	0.096	(1.613,1.988)			
σ_2	0.882	0.042	(0.8,0.965)			
ν	-.087	0.088	(-.259,0.085)			

Table 4: Parameter estimates of model (5)

	Estimate	SE	95% CI
intercept	34.06	0.105	(33.85,34.27)
preeclampsia-unrelated birth (Yes)	0.348	0.100	(.152,0.544)
parity (> 0)	0.682	0.086	(0.413,0.751)
number of fetuses (> 1)	-0.374	0.110	(-0.590,-0.158)
smoker (Yes)	-0.542	0.134	(-0.805,-0.279)
parity × history of preterm birth (Yes)	-0.189	0.075	(-0.336,-0.042)
parity × history of preterm birth (Yes)× number of fetuses	-1.330	0.361	(-2.038,-0.622)
ρ_{11}	0.469	0.179	(0.118,0.820)
ρ_{22}	0.182	0.032	(0.119, 0.245)
ρ_{12}	0.071	0.205	(-0.331,0.473)

Table 5: Results of the simulation Study

Parameter	value	Monte-Carlo mean	Monte-Carlo SE	Square root of Monte-Carlo mean of variance	95% cov. prob.
α_{01}	-3.81	-3.81	0.060	0.054	96.5
α_{02}	-2.42	-2.42	0.030	0.032	96.3
α_1	-1	-1	0.074	0.073	95.4
α_2	-0.5	-0.50	0.043	0.044	95.6
σ_1	2	2.00	0.053	0.055	94.8
σ_2	1.5	1.50	0.034	0.035	94.6
ν	0.5	0.50	0.030	0.032	95.9
β_0^*	34.68	34.68	0.027	0.028	96.3
β_1^*	0.11	0.11	0.048	0.048	93.7
θ^*	-0.12	-0.12	0.067	0.069	95.9
γ	-.0070	-.0070	0.14	0.13	91.9
ρ	0.30	0.30	0.022	0.022	93.2