Published in final edited form as:

Stat Med. 2018 May 30; 37(12): 1973–1985. doi:10.1002/sim.7624.

Estimating Recurrence and Incidence of Preterm Birth Subject to Measurement Error in Gestational Age: A Hidden Markov Modeling Approach

Paul S. Albert

Biostatistics Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute, Rockville, MD 20852

Abstract

Prediction of preterm birth as well as characterizing the etiological factors affecting both the recurrence and incidence of pre-term birth (defined as gestational age at birth < 37 weeks) are important problems in obstetrics. The NICHD consecutive pregnancy study recently examined this question by collecting data on a cohort of women with at least two pregnancies over a fixed time interval. Unfortunately, measurement error due to the dating of conception may induce sizable error in computing gestational age at birth. This article proposes a flexible approach that accounts for measurement error in gestational age when making inference. The proposed approach is a hidden Markov model that accounts for measurement error in gestational age by exploiting the relationship between gestational age at birth and birth weight. We initially model the measurement error as being normally distributed, followed by a mixture of normals that has been proposed based on biological considerations. We examine the asymptotic bias of the proposed approach when measurement error is ignored and also compare the efficiency of this approach to a simpler hidden Markov model formulation where only gestational age and not birth weight is incorporated. The proposed model is compared with alternative models for estimating important covariate effects on the risk of subsequent preterm birth using a unique set of data from the NICHD consecutive pregnancy study.

Keywords

Hidden Markov model; Longitudinal data analysis; Measurement error

1. Introduction

The NICHD Consecutive Pregnancy Study (CPS) was conducted in the hopes of identifying predictors of poor pregnancy outcomes in successive pregnancies. In this study, investigators obtained records from a large network of hospitals in Utah over the time interval 2002–2010 ([1]). Successive pregnancies were linked together and the goal was to develop predictors of poor pregnancy outcomes from a women's pregnancy-specific covariates and prior pregnancy history. Preterm birth is an important pregnancy outcome whose consequences

*Correspondence to: Paul S. Albert, Biostatistics Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute, Rockville, MD 20852.

constitute major health problems worldwide ([2]). Identifying risk factors that affect the recurrence of preterm birth is important for managing high risk pregnancies. Grantz et al. ([3]) used longitudinal transition models to compare the effect of various factors on both the incidence and prevalence of preterm birth in consecutive pregnancies. In these analyses, preterm birth was determined for each pregnancy by thresholding gestational age at 37 weeks, where gestational age is determined by either the period of the last menstrual period (LMP) or ultrasound during the first ten weeks of pregnancy. Both these measures for dating conception are known to contain measurement error ([4]; [5]) which is impossible to directly access in most practical situations. This paper proposes a latent modeling approach that accounts for measurement error in gestational age by exploiting the relationship between gestational age and birth weight.

In this article, We focus on the 27077 women who are nulliparous (experiencing their first live birth) in the 2002–2010 time interval. Without accounting for measurement error, we can simply calculate the proportion of women who have a gestational age at or below 37 weeks. Calculated this way, the preterm birth rates are 7.7% and 7.0% for the first and second pregnancy in the CPS study, respectively. Further, the preterm birth rate is 25.7% for the second pregnancy given that the first pregnancy resulted in a preterm birth. These simple summary statistics show the importance of estimating transition probabilities across consecutive pregnancies, but clearly do not account for measurement error in gestational age determination.

We had very few women missing birth weight or gestational age with 27054 subjects having complete birth weight and gestational age information on the first two pregnancies used in the analyses. Further, only 19 women had missing smoking data at the start of the second pregnancy, an important covariate in the analysis.

Figure 1 shows the relationship between gestational age (calculated by ultrasound or using the LMP if ultrasound was not available) and log-transformed birth weight for the first two pregnancies for each women in the study. The log transformation was chosen for birth weight since this outcome is approximately normally distributed with constant variance on this scale. In both plots, the solid line denotes the fitted value for a cubic polynomial fit with least-squares regression, while the vertical dotted line denotes the cutoff for preterm birth at 37 weeks. The figure suggests the potential for birth weight to provide information about the measurement error in gestational age. For example, slightly below the cutoff, there are a number of neonates that have a birth weight that is unusually high for a preterm birth. Likewise, slightly above the cutoff, there are a number of neonates that have a birth weight that is unusually low, raising the possibility that their gestational ages are in error, resulting in a misclassification with regard to preterm birth. The approach taken in this paper is to exploit this information in parameter estimation.

The proposed methodology is based on formulating a hidden Markov model ([6]) where the outcome is an indicator of whether the true gestational age is below 37 weeks at each pregnancy, and interest is on characterizing transitions of this outcome across consecutive pregnancies. [7], [8], and [9] dichotomize a single continuous biomarker and propose a hidden Markov model to estimate the transition probabilities accounting for measurement

error for this dichotomized biomarker. There has been extensions of hidden Markov models to allow for multivariate responses at each follow-up time ([10] and [11], for example). Ip et al. [12] proposed a hidden Markov model for estimating regression parameters that characterize the transition probabilities and measurement error across ordinal states in a study of disability in older adults. The approach in the current paper is to use the additional information inherent in the birth weight versus gestational age relationship to precisely estimate the measurement error of gestational age and the regression parameters characterizing the transition probabilities, which is our primary interest. This is in contrast to a mixture model proposed by [13] that uses a three component mixture model in developing a birth normality index that gauges whether pairs of birth weight and gestational age are consistent with premature, at-risk, or healthy births.

We propose both a simple Gaussian distribution as well as a flexible mixture of three Gaussian distributions for the measurement error in gestational age, and use the additional information on birth weight to estimate this measurement error. Although this article proposes this new methodology in the context of characterizing preterm birth in consecutive pregnancy, the statistical approach can be applied more generally. For example, the methodology will be useful in any situation where the outcome is determined by dichotomizing a continuous measurement measured with error, understanding the transition process of being below/above a stated threshold is of interest, and we can model the relationship between the continuous measurement of interest and an auxiliary variable to improve the measurement error estimation. In Section 2, we propose a hidden Markov model for the bivariate response of gestational age and birth weight, and discuss parameter estimation for this model. We also formulate a simpler hidden Markov model that incorporates gestational age but not birth weight to show the efficiency gains obtained by incorporating the relationship between birth weight and gestational age. In Section 3, we examine the asymptotic bias in traditional methods that ignore the measurement error in gestational age. Further, we evaluate the asymptotic efficiency gain in incorporating birth weight into the analysis as compared with only using gestational age in the hidden Markov modeling framework. In Section 4, we show the simulation study results that examine the importance of accounting for the measurement error in gestational age and for exploiting information from birth weight for estimation. We use the methodology to analyze data from the CPS and show how failure to account for measurement error in gestational age may lead to biased estimation in Section 5. A discussion follows in Section 6.

2. Model Formulation and Parameter Estimation

Let Y_{ij} , T_{ij} , and W_{ij} be the observed gestational age, true gestational age, and appropriately transformed birth weight, respectively, for the j th pregnancy on the i th subject, where $i = 1, 2, \dots, I$, $j = 1, 2, \dots, n_i$, I is the number of subjects, and n_i is the number of followup pregnancies on the i th subject. Further, denote $Y_i = (Y_{i1}, Y_{i2}, \dots, Y_{in_i})$ and $W_i = (W_{i1}, W_{i2}, \dots, W_{in_i})$. Define $Y_{ij} = T_{ij} + \varepsilon_{ij}$ where ε_{ij} is the measurement error for gestational age, and T_{ij} is normally distributed with mean μ_T and variance σ_T^2 . We initially model the measurement error for gestational age ε_{ij} as Gaussian with mean zero and variance where σ_ε . We log-transform birth weight and assume that $W_{ij}/T_{ij} = t$ is normally distributed with

mean $\omega_0 + \omega_1 t + \omega_2 t^2 + \omega_3 t^3$ and variance σ_w^2 . The cubic polynomial representation of the mean structure and constant variance on the log birthweight characterize the data well (Figure 1). Other factors may affect birth weight, requiring us to incorporate covariates as additive or interactive effects with t to adequately characterize the relationship between gestational age and birth weight.

Define $Z_{ij} = 1$ as $T_{ic} < c$ where $c = 37$ weeks is the traditional definition of preterm birth. The scientific interest is on predicting preterm birth in consecutive pregnancies. Specifically, we are interested in examining whether women-specific or pregnancy-specific covariates differentially affect incidence and recurrence. Thus, we can formulate the question with the following transition model ([14]),

$$\text{logit}P(Z_{ij} = 1 \mid Z_{i(j-1)}) = X'_{ij}\beta + Z_{i(j-1)}X'_{ij}\eta, \quad (1)$$

where covariates X_{ij} influence both the incidence and recurrence of preterm birth. For example, β reflects the effect of covariates on the risk of preterm birth for a pregnancy where the previous pregnancy was not preterm (i.e., incidence). Likewise, $\beta + \eta$ reflects the effect of covariates on the risk of a preterm birth for a pregnancy where the previous pregnancy was a preterm birth (i.e., recurrence). Without covariates, $\text{logit}P(Z_{ij} = 1 \mid Z_{i(j-1)} = z) = \beta_1 + \eta_1 z$, where the incidence is $P_{01} = \exp(\beta_1) / (1 + \exp(\beta_1))$ and the recurrence is $P_{11} = \exp(\beta_1 + \eta_1) / (1 + \exp(\beta_1 + \eta_1))$. With a Bayesian estimation approach, computation may be simpler with a probit rather than a logit link function in (1). However, for the maximum-likelihood approach that we are proposing, the computation is similar for either link function, and parameters are more interpretable with the logit link.

Denote γ as the probability of a preterm in the first pregnancy (initial probability), $\gamma = P(Z_{i1} = 1)$. In order to parameterize according to (3.1), we write the individual's contribution to the likelihood, $P(Y_i, W_i)$, as

$$\begin{aligned} P(Y_i = y_i, W_i = w_i) &= \sum_{i_1=0}^1 \sum_{i_2=0}^1 \cdots \sum_{i_{n_i}=0}^1 \{\gamma^{i_1} (1-\gamma)^{(1-i_1)} \prod_{j=2}^{n_i} P(Z_{ij} = i_j \mid Z_{i(j-1)} = i_{j-1}) \\ &\quad \} \times \left\{ \prod_{j=1}^{n_i} f_{Y,W \mid Z=1}(y_{ij}, w_{ij})^{i_j} f_{Y,W \mid Z=0}(y_{ij}, w_{ij})^{1-i_j} \right\}, \end{aligned} \quad (2)$$

where i_j is a binary variable (0/1) that is equal to 1 when $Z_{ij} = 1$ and zero, otherwise. Further, under the assumption that Y and W are independent conditional on the true gestational age at birth t , $f_{Y,W \mid Z=1}(y, w)$ and $f_{Y,W \mid Z=0}(y, w)$ can be written as

$$f_{Y, W | Z=1}(y, w) = A_1 \int_{-\infty}^c \phi(y, t, \sigma_\varepsilon^2) \phi(w, \omega_0 + \omega_1 t + \omega_2 t^2 + \omega_3 t^3, \sigma_W^2) \phi(t, \mu_T, \sigma_T^2) dt \quad (3)$$

and

$$f_{Y, W | Z=0}(y, w) = A_2 \int_c^\infty \phi(y, t, \sigma_\varepsilon^2) \phi(w, \omega_0 + \omega_1 t + \omega_2 t^2 + \omega_3 t^3, \sigma_W^2) \phi(t, \mu_T, \sigma_T^2) dt \quad (4)$$

where

$$A_1 = \frac{1}{\Phi(c) \sigma_\varepsilon \sigma_W \sigma_T}, A_2 = \frac{1}{(1 - \Phi(c)) \sigma_\varepsilon \sigma_W \sigma_T}, \quad (5)$$

$\phi(x, \mu, \sigma_2)$ is a normal density with mean μ and variance σ_2 , and $\Phi(x)$ is the cumulative standard normal distribution evaluated at x . These calculations assume that Y and W are statistically independent given t .

Model (3)–(5) correspond to the case where measurement error is Gaussian. However, there is a suggestion in the reproductive epidemiology literature that measurement error for gestational age may be trimodal, where one group is overestimated, another is underestimated, and the third has a mode which is centered at zero ([4]). Specifically, $\varepsilon_{ij} \sim p_1 N(\mu_1, \sigma_\varepsilon^2) + p_2 N(\mu_2, \sigma_\varepsilon^2) + (1 - p_1 + p_2) N(0, \sigma_\varepsilon^2)$, where $\mu_2 = -p_1 \mu_1 / p_2$ since $E(\varepsilon_{ij}) = 0$. Under this formulation, p_1 is the probability of being in the group that overestimates gestational age by an average of μ_1 weeks (i.e., earlier bleeding is characterized as pregnancy), p_2 is the probability of being in the group that underestimates gestational age by an average of μ_2 weeks (e.g., miss a cycle based on recall bias in reporting the LMP), and $1 - p_1 - p_2$ is the probability of being in the group that, on average, correctly estimates gestational age. In each group, additional measurement error variation is introduced by allowing for normal variation with mean zero and variation σ_ε^2 . Even if we are not completely comfortable interpreting this three group mixture model from a biological perspective, the formulation provides a flexible way to represent the measurement error distribution to obtain unbiased estimates of the transition parameters, the key parameters of scientific interest.

For this more general model formulation,

$$f_{Y,W|Z=1}(y,w) = A_1 \int_{-\infty}^c [p_1 \phi(y, t + \mu_1, \sigma_\varepsilon^2) + p_2 \phi(y, t + \mu_2, \sigma_\varepsilon^2) + (1 - p_1 + p_2) \phi(y, t, \sigma_\varepsilon^2)] \\ \times \phi(w, \omega_0 + \omega_1 t + \omega_2 t^2 + \omega_3 t^3, \sigma_W^2) \phi(t, \mu_T, \sigma_T^2) dt$$

(6)

and

$$f_{Y,W|Z=0}(y,w) = A_2 \int_c^\infty [p_1 \phi(y, t + \mu_1, \sigma_\varepsilon^2) + p_2 \phi(y, t + \mu_2, \sigma_\varepsilon^2) + (1 - p_1 + p_2) \phi(y, t, \sigma_\varepsilon^2)] \\ \times \phi(w, \omega_0 + \omega_1 t + \omega_2 t^2 + \omega_3 t^3, \sigma_W^2) \phi(t, \mu_T, \sigma_T^2) dt.$$

(7)

Integrals (3)–(4) and (6)–(7) are approximated using a simple trapezoidal rule. Maximum-likelihood follows by maximizing the log-likelihood $L = \sum_{i=1}^I \log P(Y_i, W_i)$, where $P(Y_i, W_i)$ is given by (2). For numerical stability, we model σ_W , σ_T and σ_ε on the log-scale and μ_T and γ on the logit scale. Further, for the three group mixture of Gaussian distributions, we model p_1 and p_2 with a polychotomous logit transformation. Specifically, $p_1 = \exp(\eta_1) / [1 + \exp(\xi_1) + \exp(\xi_2)]$, $p_2 = \exp(\xi_2) / [1 + \exp(\xi_1) + \exp(\xi_2)]$, where $1 - p_1 - p_2 = 1 / [1 + \exp(\eta_1) + \exp(\xi_2)]$.

To show the advantages of incorporating birth weight into the modeling strategy, we note that the hidden Markov model can be formulated using only gestational age and not birth weight. When only gestational age is incorporated, each individual's contribution to the likelihood is $P(Y_i = y)$, which can be evaluated using equation (2) with $f_{Y,W|Z=1}$ and $f_{Y,W|Z=0}$ replaced by $f_{Y|Z=1}$ and $f_{Y|Z=0}$, respectively, where

$$f_{Y|Z=1}(y) = B_1 \int_{-\infty}^c \phi(y, t, \sigma_\varepsilon^2) \phi(t, \mu_T, \sigma_T^2) dt \quad (8)$$

and

$$f_{Y|Z=0}(y) = B_2 \int_c^{\infty} \phi(y, t, \sigma_e^2) \phi(t, \mu_T, \sigma_T^2) dt. \quad (9)$$

where $B_1 = 1/(\Phi(c)\sigma_e\sigma_T)$ and $B_1 = 1/(1 - \Phi(c))\sigma_e\sigma_T$.

Maximum-likelihood was implemented by using the optim function in R (version 3.2.2) with the limited-memory Broyden-Fletcher-Goldfarb-Shanno (L-BFGS) algorithm [19]. The algorithm numerically approximates the Hessian matrix that is used to estimate variances of parameter estimates. The R-code for parameter estimation is available from the author upon request.

3. Asymptotic Bias and Efficiency

In this section, we address two important practical questions from a theoretical perspective. First, what is the effect of estimating incidence and recurrence when we ignore the measurement error in gestational age. Second, as mentioned at the end of Section 3, we can obtain properly corrected estimates of incidence and prevalence when we ignore the birth weight distribution and only incorporate gestational age into the modeling. However, one might expect that such an approach would be inefficient relative to the incorporation of birth weight. We examine this issue by evaluating the asymptotic relative efficiency for estimation that jointly models gestational age and birth weight as compared to the approach that only uses gestational age.

The asymptotic bias and variance calculations are based on two pregnancies ($n_i = 2$, for all i). To compute the asymptotic bias in estimating incidence and prevalence, we first note that the estimated incidence and recurrence at the second pregnancy when we ignore the measurement error can be estimated as

$$\widehat{Incid} = \frac{\sum_{i=1}^I I(Y_{i1} > c) I(Y_{i2} \leq c))}{\sum_{i=1}^I E(I(Y_{i1} > c))} \text{ and } \widehat{Recur} = \frac{\sum_{i=1}^I I(Y_{i1} \leq c) I(Y_{i2} \leq c))}{\sum_{i=1}^I I(Y_{i1} \leq c)} \quad (10)$$

The estimators $\widehat{Incid}$ and $\widehat{Recur}$ will converge to P_{01}^* and $1 - P_{10}^*$ in large samples, where

$$P_{01}^* = \frac{E(I(Y_{i1} > c) I(Y_{i2} \leq c)))}{E(I(Y_{i1} > c))} \text{ and } P_{10}^* = \frac{E(I(Y_{i1} \leq c) I(Y_{i2} > c)))}{E(I(Y_{i1} \leq c))}. \quad (11)$$

Without loss of generality, we will show expressions for $E(I(Y_{i1} > c)I(Y_{i2} \leq c))$ and $E(I(Y_{i1} > c))$ for computing P_{01}^* . The expressions corresponding to P_{10}^* can be similarly derived. The numerator and denominator of P_{01}^* can be expressed as

$$\begin{aligned}
 E[I(Y_{i1} > c)I(Y_{i2} \leq c)] &= \sum_{i_1=0}^1 \sum_{i_2=0}^1 \gamma^{i_1} (1-\gamma)^{1-i_2} P_{11}^{i_1 i_2} P_{00}^{(1-i_1)(1-i_2)} P_{01}^{(1-i_1)i_2} P_{10}^{i_1(1-i_2)} \times P \\
 &\quad (Y_{i1} > c \mid Z_{i1} = i_1) P(Y_{i2} \leq c \mid Z_{i2} = i_2) \\
 E[I(Y_{i1} > c)] &= \sum_{i_1=0}^1 \gamma^{i_1} (1-\gamma)^{1-i_1} P(Y_{i1} > c \mid Z_{i1} = i_1).
 \end{aligned}
 \tag{12}$$

Finally, to evaluate P_{01}^* we need

$$\begin{aligned}
 P(Y_{ij} > c \mid Z_{ij} = 1) &= \frac{\Phi_{\mu, \sigma_T^2 + \sigma_\epsilon^2}^{(c)} - \Phi_{(\mu, \mu_T)'}^{(c, c)} \Sigma^{(c, c)}}{\Phi_{\mu, \sigma_T^2}^{(c)}}, P(Y_{ij} > c \mid Z_{i1} = 0), P(Y_{ij} \leq c \mid Z_{i1} = 0) = 1 \\
 &= \frac{1 - \Phi_{\mu, \sigma_T^2}^{(c)} - \Phi_{\mu, \sigma_T^2 + \sigma_\epsilon^2}^{(c)} + \Phi_{(\mu, \mu_T)'}^{(c, c)} \Sigma^{(c, c)}}{1 - \Phi_{\mu, \sigma_T^2}^{(c)}}
 \end{aligned}$$

$- P(Y_{ij} > c \mid Z_{i1} = 0)$, and $P(Y_{ij} \leq c \mid Z_{i1} = 1) = 1 - P(Y_{ij} > c \mid Z_{i1} = 1)$, where

$$\Sigma = \begin{pmatrix} \sigma_T^2 + \sigma_\epsilon^2 & \sigma_T^2 \\ \sigma_T^2 & \sigma_T^2 + \sigma_\epsilon^2 \end{pmatrix}, \Phi_{\mu, \sigma^2}(c) \text{ is a cumulative normal distribution with mean } \mu \text{ and}$$

variance σ^2 , evaluated at c , and $\Phi_{(\mu_1, \mu_2)'}^{(c, c)} \Sigma$ is a cumulative bivariate normal distribution with mean $(\mu_1, \mu_2)'$ and variance Σ .

Using the above expressions, we are able to compute the relative bias for incidence and recurrence by evaluating $(P_{01}^* - P_{01})/P_{01}$ and $(P_{10} - P_{10}^*)/(1 - P_{10})$, respectively. The bias is computed for an incidence of 2.4% and a recurrence of 33% and with other parameters closely approximating the ones that are estimated from the CPS dataset (See Table 2 which is discussed further in Section 4). Figure 2 shows the relative bias for incidence and recurrence as a function of the standard deviation of the measurement error. Note that since we are computing asymptotic bias (bias as I gets large), no sample size I is specified. The bias is large even for relatively small measurement error. For example, the measurement error standard deviation estimated from the CPS dataset is 0.60 weeks, corresponding to a relative bias of 400% and -30%, respectively. These calculations demonstrate the importance of accounting for the measurement error in this problem. The intuition for why measurement error results in biased estimation is that although the measurement error for

gestational age is assumed to be symmetric, it is being applied and averaged across non-linear function.

We also evaluated the asymptotic relative efficiency of incorporating gestational age and birth weight (equations 1–5) as compared with only incorporating gestational age in the hidden Markov modeling formulation (equations 1,8–9). For each formulation, the asymptotic variances of β_1 and η_1 is obtained by evaluating the expected value of the inverse of the hessian matrix. We computed the expected value by Monte-Carlo sampling where we simulated 100,000 individual realizations, compute the hessian matrix for each realization and then averaging them. We computed the asymptotic relative efficiency for both β_1 and η_1 for parameter values close to the ones that are estimated from the CPS dataset in Section 4 except we varied the values of σ_W , the residual variance of the birth weight regression. Figure 3 shows plots of the relative efficiency of β_1 and η_1 as a function of reasonable values of σ_W . We see that there is large relative efficiency in incorporating birth weight when the birth weight variance is small. At $\sigma_W = 0.05$ the relative efficiency for both parameters is close to 2.5, demonstrating a very large gain in incorporating birth weight into the analysis. When σ_W is large, the relative efficiency is near 1, demonstrating that there is little benefit in incorporating birth weight into the analysis. For the CPS analysis, σ_W is estimated close to 0.11, demonstrating a nearly 50% efficiency gain in incorporating birth weight into the analysis over simply using gestational age.

4. Estimating the incidence and recurrence of preterm birth in the NICHD Consecutive Pregnancy Study

We limited the analysis to the first two pregnancies on each participant since a majority of women have only two pregnancies over the 10 year window (of the 51,066 women in the study, 39,961 (78%) had only two pregnancies). Further, we limited our analysis to those women who were nulliparous (no prior live births) at the first pregnancy in the window (this included 27,077 out of the 51,066) and to women who had their first pregnancy within the first year of the study (2002–2003). We also excluded the very small percentage of women who had missing birth weight or covariate information. The sample size for this analysis was $I = 5296$. Thus, our inferences can be generalized to all women in the population who have at least two pregnancies over a seven seven year follow-up period. The primary objective of this analysis is to estimate the incidence and recurrence of preterm birth for the second pregnancy and to examine the differential effect of covariates on these proportions. We fit transition regression models where we examine whether risk factors affect incidence differently than recurrence of preterm birth for the second pregnancy. Four approaches are considered, (i) a logistic regression model that ignores the measurement error in gestational age, (ii) the proposed hidden Markov model where measurement error of gestational age is assumed to be Gaussian distributed, (iii) a hidden Markov model that only uses gestational age and not birth weight, and assumes Gaussian measurement error, and (ii) the proposed model where measurement error for gestational age is assumed to be a mixture of three Gaussian distributions.

Grantz [3] investigated whether demographic factors differed between the incidence and prevalence of preterm birth in the second pregnancy (interaction between factors and first-term preterm birth status) using CPS data. The authors found a statistically significant two-way interaction with both smoking and alcohol-use with first pregnancy preterm birth status. Such an interaction was not found for the gap-time between the first and second pregnancy and first pregnancy preterm birth. The current analysis focuses on the smoking by preterm birth status accounting for measurement error in the gestational age determination. Table 1 shows maximum-likelihood estimates using logistic regression when we ignore the measurement error. For numerical stability, we re-scaled time by dividing time in weeks by 42. First, we fit a logistic regression where the probability of a women having a preterm birth (gestational age < 37 weeks) on the logit scale is modeled as a linear function of whether the first pregnancy was preterm. The results show that the odds of having a preterm birth in the second pregnancy for a women who had a preterm birth in the first pregnancy is $\exp(1.92) = 6.8$ times that for a women who did not have a preterm birth at the first pregnancy. Second, in order to examine whether the etiological effect of smoking was different for incidence versus recurrence of a preterm birth, we include both a main effect of smoking as well as an interaction between first pregnancy preterm birth and smoking. The results suggest that there is a strong first pregnancy preterm birth by smoking interaction ($\hat{\beta}_3 = -1.90$ (SE=0.69)) suggesting that smoking effects incidence more than recurrence. Specifically, the odds ratio of a second pregnancy preterm birth for smokers versus non-smokers is $\exp(2.02) = 7.5$ for incidence and $\exp(2.02 - 1.897) = 1.2$ for recurrence, respectively.

Table 2 shows maximum-likelihood estimates obtained using (1) to (5) which assumes Gaussian measurement error. Estimation is presented for both the case of no covariates as well as the model that includes a smoking by first-pregnancy preterm birth interaction. The model presented on the left side of Table 2 shows estimates for the case of no covariates. Consistent with the attenuation observed in the bias calculations of Section 2, the estimates of β_1 and η_1 are larger in magnitude than the logistic regression that does not account for measurement error (Table 1). With the proposed model, the odds of having a preterm birth on the second pregnancy for a women who had a first pregnancy preterm birth is $\exp(3.03) = 20.70$ times that for a women who did not have a first pregnancy preterm birth. The estimate of measurement error variance for gestational age is $\exp(-4.08) = 0.017$, which on a weekly scale results in a standard deviation of $0.017 * 42 = 0.714$ or 5 days. This is consistent with measurement error standard deviations of 5–7 days that are reported in the literature for estimating gestational age ([15]). Estimates of μ_T and σ_T^2 , that are on the logit and log scales, respectively, show a mean gestational age at delivery of 0.90 and 0.054, respectively. On a weekly scale, this translates to a mean of $0.90 * 42 = 37.8$ weeks (SD= $0.054 * 42 = 2.30$ weeks). The probability of being truly preterm at the first pregnancy is -2.87 on the logit scale, or 0.054. In this analysis, we included a smoking effect in modeling the relationship between gestational age and birth weight, where smokers resulted in a slightly lower log-transformed birthweight ($\omega_4 = -0.057$).

A hidden Markov model using only gestational age and not birth weight was also fit to these data. Parameter estimates were similar to those presented in Table 2, but as expected, there

was efficiency loss. For a model without covariates, the parameters β_1 and η_1 were estimated as -3.70 ($SE=0.12$) and 3.06 ($SE=0.21$), respectively, showing a 19% and 14% efficiency loss relative to the model that exploits the relationship between gestational age and birth weight.

The right side of Table 2 provides the maximum-likelihood estimates from the model that includes a smoking (at beginning of second pregnancy) by first-pregnancy preterm birth interaction. Again, compared to the logistic regression that examines this interaction, the effects are enhanced in magnitude for the model that incorporates measurement error. There is a strong first pregnancy preterm birth by smoking interaction ($\hat{\beta}_3 = -3.08$ ($SE=1.23$)) with the odds ratio of a second pregnancy being preterm among smokers versus non-smokers is $\exp(3.23) = 25$ for incidence and $\exp(3.23 - 3.08) = 1.16$ for recurrence.

Table 3 shows maximum-likelihood estimates for the three group mixture model error structure described by (6) to (7). The three group mixture model fits substantially better than the model that allows only for a Gaussian measurement error. Specifically, or the model without covariates, the AIC for the Gaussian and the three-group mixture model was $-52,732$ and $-55,733$, respectively, providing evidence for the three-group mixture model. Similarly, the AIC was smaller for the three-group mixture model than the Gaussian model when a smoking by first-pregnancy-preterm birth interaction is introduced (See Tables 2 and 3). Estimates of ξ_1 and ξ_2 are polychotomous transformed values of p_1 and p_2 . Specifically, estimates of p_1 and p_2 are $\exp(-3.75)/[1+\exp(-3.75)+\exp(-0.029)]=0.012$ and $\exp(-0.029)/[1+\exp(-3.75)+\exp(-0.029)]=0.487$, respectively. interpretation of the three group mixture model is that there is a small probability of being in a group that overestimates the gestational age and a large probability of being in the group whose mean underestimates the gestational age. The mean overestimation is $\exp(-1.45)*42=9.87$ wks for group 1 and the mean underestimation is 0.243 wks. The first pregnancy preterm birth by smoking interaction was highly significant with this model ($\hat{\beta}_3 = -2.34$ ($SE=9.775$)). With this mixture model, the odds ratio of a second pregnancy being preterm among smokers versus non-smokers being $\exp(3.04) = 21$ for incidence and $\exp(3.04 - 2.34) = 2.10$ for recurrence. We fit an additional model where mother's age and the presence of chronic hypertension at the start of pregnancy are added as covariates in the regression model of W_{ij}/T_{ij} . The results showed that mother's age and chronic hypertension both increased birth weight with only minor changes to estimates of β_1 , η_1 , β_2 , and η_2 (data not shown). In general, we found that gestational age at birth was the major predictor of birth weight and the additional contribution of other potential predictors added little to the estimation of the regression coefficients β_1 , η_1 , β_2 , and η_2 .

5. Simulations

We examine the performance of the modeling approaches that incorporate the relationship between gestational age and birth weight under correct and misspecified models. Table 4 shows simulation results when the measurement error is normally distributed with parameters similar to those estimated in our application. One-thousand simulated datasets with $I=5296$, the same number as in the CPS analysis, were generated and analyzed. We fit both the proposed models and an ordinary logistic regression model that does not take into

account the measurement error in the continuous biomarker. As in the analysis of the CPS data, we rescaled time by dividing the number of weeks by 42. Under the correctly specified model that assumes a Gaussian distributed measurement error, parameter estimation is nearly unbiased. In addition, for all but the polynomial terms for the regression relating gestational age to birth weight, the average asymptotic standard errors are close to the Monte-Carlo estimates, suggesting the variance estimation for important parameter estimates performs well. The variances of the polynomial terms are poorly estimated, primarily because these estimates are very highly correlated with each other. This is inconsequential since the relationship between gestational age and birth weight is not of direct interest. Rather, as long as this relationship is accurately specified, the statistical properties for inferences about the transition probabilities, the parameters of interest, are good.

Table 4 also shows the bias in estimation when the measurement error in the continuous marker is not accounted for when using logistic regression. As the asymptotic bias calculations indicated, β_1 and η_1 are severely attenuated when the measurement error in classification is not appropriately accounted for.

Table 5 shows the importance of correctly specifying the measurement error distribution in estimation. Specifically, data are simulated under a three group mixture model for the measurement error while estimation is performed assuming a Gaussian distribution for the measurement error. We present three scenarios with an increase proportion of women who overestimate gestational age by 4 weeks and who underestimate gestational age by 4 weeks. The results show that there is increased bias in estimating β_1 and η_1 with increased departure from a normal distribution (increased proportion who both over and underestimate gestational age). Interestingly, the bias is in the direction of anti-attenuation rather than the attenuation bias we saw when we did not consider measurement error by using logistic regression.

6. Discussion

We propose new methodology for predicting the occurrence of preterm birth in consecutive pregnancies that accounts for the measurement error in gestational age. It is well known in the perinatal epidemiology that gestational age is measured with error, and in some analyses ignoring this fact can lead to substantial bias. In this paper, we show the importance of properly accounting for this measurement error in transition models. Analyses with the new methodology confirmed a smoking by first-pregnancy preterm birth interaction found in [3] based on logistic regression. However, with appropriate modeling of the measurement error, we found that the odds ratio for smoking on preterm birth was substantially larger for incidence (when the first pregnancy was not preterm) as compared with logistic regression.

Hidden Markov models provide a natural framework for accounting for this measurement error (See [7], [8], and [9], for example). Our methodology extends this work by leveraging information on another variable, birth weight, in order to obtain more information about the form of the measurement error. We show through asymptotic efficiency calculations that incorporating birth weight results in large efficiency gains when the residual error in birth weight is small to moderate as in the CPS. Further, we show through simulation and analysis

the importance of properly modeling the measurement error when the focus is on estimating the parameters of a transition model.

This methodology was applied to examine preterm birth in consecutive pregnancies, but could be applied more generally to other poor pregnancy outcomes such as large- or small-for gestational age, where a similar question is of interest. One of the major unsolved problems in obstetrics has been establishing good predictors of preterm birth. Although there are numerous reasons for this difficulty, one reason may be the misclassification of preterm birth which attenuates logistic regression estimates as well as the accuracy of prediction using standard methods. The proposed models may help in this regard.

We did not consider the situation of missing outcome (birthweight and the associated gestational age) or covariate data since the amount of such missing data was very small in the CPS data. In the CPS analysis, we simply deleted those subjects with missing data. However, in situations with a more sizable amount of missing data, multiple imputation can be used under a missing at random data mechanism assumption.

Although the major contribution of this article was to develop a model for characterizing the transition process of preterm birth across consecutive pregnancies that corrects for measurement error in gestational age estimation, this modeling strategy can be applied in other applications. The methodology can be applied in longitudinal data analysis where the outcome is obtained by dichotomizing a continuous marker that is subject to measurement error. Further, our work shows that there is substantial efficiency gain in exploiting a functional relationship between the dichotomized biomarker and another variable in order to more effectively account for measurement error. An example might include dichotomized CD4 counts and exploiting the functional relationship between CD4 and white cell counts in characterizing the transition process in repeated measurements of HIV infected subjects.

The focus of this work was on modeling a women's first two pregnancies with a maximum gap between pregnancies of 7 years. This was done since our primary interest was on examining differential etiological effects between the incidence and recurrence of preterm birth in a women's second pregnancy. Incorporating all a woman's pregnancies over the time interval is challenging since there may be selection bias due to only 12% of women had having three or more pregnancies in the study window. Chaurasia et al. [18] proposed a pattern mixture model that accounts for this type of selection bias in estimating a transition model using all of the consecutive pregnancies in the CPS. However, they did not consider measurement error in gestational age determination. Future research will focus on incorporating such measurement error in models that incorporate this type of design bias. Since only the first two pregnancies were considered, the first-order Markov model is natural model choice. Women with two or more pregnancies would need to be considered in order to investigate heterogeneity in transition probabilities across women.

Acknowledgments

This study utilized the high-performance computational capabilities of Biowulf Linux cluster at the National Institutes of Health, Bethesda, MD (<http://biowulf.nih.gov>). The research was supported by the Intramural Research Programs of the National Cancer Institute and the National Institutes of Health *Eunice Kennedy Shriver* National Institute of Child Health and Human Development.

References

1. Laughon SK, Albert PS, Leishear K, Mendola P. The NICHD Consecutive Pregnancy Study: recurrent preterm delivery by subtype. *American Journal of Obstetrics and Gynecology*. 2014; 210:e1–8. [PubMed: 24184183]
2. Behrman, RE., Butler, AS. *Preterm Birth: Causes, Consequences, and Prevention*. The National Academy Press; Washington DC: 2007.
3. Grantz KL, Hinkle SN, Mendola P, Sjaarda LA, Leishear K, Albert PS. Differences in risk factors for recurrent versus incident preterm delivery. *American Journal of Epidemiology*. 2015; 182:157–162. [PubMed: 26033931]
4. Gjessing HV, Skjaerven R, Wilcox AJ. Errors in gestational age: evidence of bleeding in pregnancy. *American Journal of Public Health*. 1999; 89:213–218. [PubMed: 9949752]
5. Hoffman CS, Messer LC, Mendola P, Savitz DA, Herring AH, Hartmann KE. Comparison of gestational age at birth based on last menstrual period and ultrasound during the first trimester. *Pediatric and Perinatal Epidemiology*. 2008; 22:587–596. [PubMed: 19000297]
6. Zucchini, W., MacDonald, IL. *Hidden Markov models for Time Series: An Introduction Using R*. Chapman & Hall; Boca Raton, Florida: 2009.
7. Satten GA, Longini IM. Markov chains with measurement error: estimating the True Course of a marker of the progression of human immunodeficiency virus disease. *Journal of the Royal Statistical Society: Series C*. 1996; 45:275–309.
8. Albert PS. A mover-stayer model for longitudinal marker data. *Biometrics*. 1999; 55:1252–1237. [PubMed: 11315078]
9. Jackson CH, Sharples LD, Thompson SG, Duffy SW, Elisabeth C. Multistate Markov models for disease progression with classification error. *Journal of the Royal Statistical Society: Series D*. 2003; 52:193–209.
10. Scott SL, James GM, Sugar CA. Hidden Markov models for longitudinal comparisons. *Journal of the American Statistical Association*. 2005; 470:359–368.
11. Ip EH, Zhang Q, Schwartz R, Tooze J, Leng X, Han H, Williamson DA. Multi-profile hidden Markov model for mood, dietary intake, and physical activity in an intervention study of childhood obesity. *Statistics in Medicine*. 2013; 32:3314–3131. [PubMed: 23322318]
12. Ip EH, Zhang Q, Rejeski J, Harris T, Kritchevsky S. Partially ordered mixed hidden Markov model for disablement process of older adults. *Journal of the American Statistical Association*. 2013; 108:370–384. [PubMed: 24058222]
13. Whitmore GA, Zhang G, Lee ML. Constructing normalcy and discrepancy indexes for birthweight and gestational age using a threshold regression mixture model. *Biometrics*. 2012; 68:297–306. [PubMed: 21838731]
14. Zeger SL, Qaqish B. Markov regression models for time series: a Quasi-likelihood approach. *Biometrics*. 1988; 44:1019–1031. [PubMed: 3148334]
15. Salem S, Kim K, VandenHof MC. Joint SOGC/CAR policy statement on non-medical use of fetal ultrasound. *Obstetrics and Gynecology Canada*. 2014; 36:184–188.
16. Baum LE, Petrie T. Statistical inference for probabilistic functions of finite state Markov chains. *Annals of Mathematical Statistics*. 1966; 37:1554–1563.
17. Baum LE, Petrie T, Soules G, Weiss N. A maximisation technique occurring in the statistical analysis of probabilistic functions of Markov chains. *Annals of Mathematical Statistics*. 1970; 41:164–171.
18. Chaurasia A, Liu D, Albert PS. Pattern-mixture models with incomplete informative cluster size: application to a repeated pregnancy study. *Journal of the Royal Statistical Society: Series C*. 2017 In Press.
19. Liu DC, Nocedal J. On the limited memory BFGS method for large scale optimization. *Mathematical Programming*. 1989; 45:503–528.

Figure 1.

log birth weight versus gestational age for all 27,054 nulliparous women in the CPS for their first and second pregnancies (left and right panel, respectively). 27,054 out of a total of 27,077 women had complete birth weight and gestational age data on both pregnancies. The dotted vertical line in each plot reflects the cutoff for preterm birth at 37 weeks. The solid line in each plot is the fitted cubic regression curve.

Figure 2.

Asymptotic relative bias for estimating incidence and recurrence (P_{01} and P_{11}) as a function of the SD of measurement error (σ_e). The bias is evaluated under model 1 with $\gamma = 0.054$, $P_{01} = 0.024$, $P_{10} = 0.33$, $\sigma_T = 2.35$ weeks, $\mu = 38.85$ weeks,

Figure 3. Asymptotic relative efficiency of β_1 and η_1 versus σ_W . The remaining parameters were chosen as the ones estimated in Table 2.

Table 1

Transition model that ignores measurement error in gestational age at birth. We fit logistic regression models with a lagged outcome with and without a smoking by lagged outcome interaction. The model is $\text{logit}P(Z_{i2} = 1|Z_{i1} = z) = \beta_1 + \eta_1 z + \beta_2 \text{Smoking}_{i2} + \eta_2 \text{Smoking}_{i2} z$. The model parameters are from equation 1.

Parameter	No Covariates	Smoking Interaction
	Est. (SE)	Est. (SE)
β_1	-2.94 (0.065)	-3.02 (0.069)
η_1	1.92 (0.132)	2.02 (0.136)
β_2	–	1.43 (0.230)
η_2	–	-1.90 (0.691)

Table 2

Transition model that assumes Gaussian measurement error in gestational age at birth. We fit model (1) with and without a smoking by lagged outcome interaction. The model is $\text{logit}P(Z_{i2} = 1/Z_{i1} = z) = \beta_1 + \eta_1 z + \beta_2 \text{Smoking}_{i2} + \eta_2 \text{Smoking}_{i2} z$. The model parameters are from equations (1) and (3–5).

Parameter	No Covariates	Smoking Interaction
	Est. (SE)	Est. (SE)
β_1	−3.68 (0.11)	−3.86 (0.12)
η_1	3.03(0.20)	3.23 (0.20)
β_2	–	2.23 (0.27)
η_2	–	−3.08 (1.11)
ω_0	7.01 (1.25)	7.41 (1.24)
ω_1	−17.61 (4.67)	−17.58 (4.65)
ω_2	37.85(5.75)	37.79 (5.72)
ω_3	−19.44 (2.32)	−19.42 (2.31)
ω_4	–	−0.047 (0.0069)
$\log \sigma_W$	−2.17 (0.0091)	−2.17 (0.0090)
$\log \sigma_T$	−2.92 (0.0159)	−2.92 (0.0158)
$\text{logit}\mu_T$	2.18 (0.0198)	2.18 (0.0199)
$\text{logit}\gamma$	−2.87 (0.069)	−2.88 (0.069)
$\log \sigma_e$	−4.08 (0.032)	−4.08 (0.031)
AIC	−55,733	−56,276

Table 3

Transition model that incorporates a three group mixture of Gaussian distributions for the measurement error in gestational age at birth. We fit model (1) with a lagged outcome with and without a smoking by lagged outcome interaction. The model is $\text{logit}P(Z_{i2} = 1/Z_{i1} = z) = \beta_1 + \eta_1 z + \beta_2 \text{Smoking}_{i2} + \eta_2 \text{Smoking}_{i2} z$. The model parameters are from equations (1) and (6–7).

Parameter	No Covariates Est. (SE)	Smoking Interaction Est. (SE)
β_1	−4.15(0.13)	−3.36 (0.11)
η_1	2.81(0.27)	3.04 (0.18)
β_2	–	2.00 (0.27)
η_2	–	−2.34 (0.78)
ω_0	−39.54 (42.52)	−1.18 (1.32)
ω_1	40.07 (143.14)	17.13 (5.13)
ω_2	75.27(160.39)	−6.49 (6.56)
ω_3	67.95 (59.82)	−1.17 (2.77)
ω_4	–	−0.50 (0.0071)
$\log \sigma_W$	−2.31 (0.013)	−2.15 (0.0084)
$\log \sigma_T$	−3.84 (0.020)	−2.95(0.013)
$\text{logit}\mu_T$	2.57 (0.0064)	1.39 (0.018)
$\text{logit}\gamma$	−3.39 (0.092)	−2.41 (0.064)
$\log \sigma_e$	−3.38 (0.00018)	−4.33 (0.032)
$\log \mu_1$	−1.45 (0.018)	98.60 (1049)
ξ_1	−3.75 (0.14)	−100 (1049)
ξ_2	−0.029 (0.11)	1.86 (0.069)
AIC	−55,733	−56,276

Table 4

Simulation study examining the statistical properties for a misspecified logistic regression model (1) that ignores the measurement error in gestational age determination and a correctly specified model that assumes Gaussian measurement error for gestational age at birth (equations (1) along with (3–5)). Simulations are done with true parameter values set to those estimated in Section 4. The sample size was set as $I = 5296$, the same as in the CPS example. The transition model assumed does not have any covariates. Specifically, we fit $\text{logit} P(Z_2 = 1 | Z_1 = z) = \beta_1 + \eta_1 z$. The simulation was done with 1,000 simulated data sets with the Mean being the average of the parameter estimates across data sets, Mean (SE) being the average standard error across data sets, and MC(SE) being the Monte-Carlo standard error.

Parameter	True Value	Misspecified Model			Correct Model		
		Mean	Mean(SE)	MC(SE)	Mean	Mean(SE)	MC(SE)
β_1	-3.68	-2.20	0.049	0.049	-3.63	0.140	0.135
η_1	3.03	0.049	0.112	0.111	2.96	0.236	0.229
ω_0	7.40	-	-	-	7.33	0.236	0.229
ω_1	-17.61	-	-	-	-17.54	0.269	14.40
ω_2	37.85	-	-	-	37.92	0.224	15.69
ω_3	-19.44	-	-	-	-19.51	0.259	5.69
$\log \sigma_W$	-2.17	-	-	-	-2.17	0.00812	0.00861
$\log \sigma_T$	-2.92	-	-	-	-2.94	0.0163	0.0165
$\text{logit} \mu_T$	2.18	-	-	-	2.21	0.0191	0.0199
$\text{logit} \gamma$	-2.87	-	-	-	-2.84	0.0807	0.0814
$\log \sigma_e$	-4.08	-	-	-	4.06	0.0275	0.0289

Table 5

Simulation: Effect of assuming a Gaussian measurement error as in equations (3–5) when the true model incorporates a three group mixture of Gaussian distributions as in equations (6–7). We simulated data with parameters and sample size as given in Table 4 with the exception of a three group mixture of Gaussian distributions rather than a Gaussian distribution. (A) represents a slight departure from a Gaussian measurement error where 5% are over-estimated by an average of four weeks, 5% are under-estimated by an average of four weeks, and the remaining 80% have mean zero. (B) represents a moderate departure from a Gaussian measurement error where 10% are over-estimated by an average of four weeks and 10% are under-estimated by an average of four weeks. Last, (C) represents a severe departure where 20% are over-estimated by an average of four weeks and 20% are under-estimated by an average of four weeks.

Parameter	True Value	Model A	Model B	Model C
		Mean (MC SE)	Mean (MC SE)	Mean (MC SE)
β_1	-3.68	-3.82 (0.145)	-3.92 (0.164)	-4.03 (0.190)
η_1	3.03	3.20 (0.243)	3.32 (0.272)	3.51 (0.314)
ω_0	7.40	5.08 (9.57)	-5.48 (19.7)	-22.86 (25.6)
ω_1	-17.61	-7.76 (30.9)	27.71 (62.8)	84.91 (80.9)
ω_2	37.85	24.85 (33.2)	-14.60 (66.6)	-77.00 (84.8)
ω_3	-19.44	-13.94 (18.3)	0.548(23.5)	23.10 (29.6)
$\log \sigma_W$	-2.17	-2.13 (0.00836)	-2.10 (0.00816)	-2.06 (0.00981)
$\log \sigma_T$	-2.92	-2.78 (0.0177)	-2.68 (0.0167)	-2.59 (0.0155)
$\logit \mu_T$	2.18	2.15 (0.0239)	2.15 (0.0271)	2.26 (0.0312)
$\logit \gamma$	-2.87	-2.99 (0.0900)	-3.07 (0.102)	-3.15 (0.114)
$\log \sigma_e$	-4.08	-4.03 (0.0363)	-4.03 (0.0572)	-4.02 (0.0783)