Misclassification Adjustment in Threshold Models for the Effects of Subject-Specific Exposure Means and Variances

Chengxing Lu* Robert H. Lyles*

Abstract

In environmental epidemiology, researchers sometimes assume the existence of exposure thresholds above which the risk of adverse effects begins to increment. In this work, we assume exposures are measured repeatedly over time, and the research question of interest is to identify the relationship between a health-related outcome and whether or not the subject-specific mean and/or variability of the exposure exceed known thresholds. As a subject's true exposure mean and variability cannot be observed directly, misclassification typically arises in the categorization of whether these quantities are above their respective thresholds. Building off of random effects models for repeated exposure measurements and assuming balanced data, methods based on the well-known regression calibration and matrix methods are demonstrated for the case of the mean exposure only. For unbalanced data, and to incorporate categorizations based on both the exposure mean and variance, a maximum likelihood approach is introduced. Simulation results and a real study example from the Mount Sinai Study of Women Office Workers (MSSWOW) are presented to demonstrate the performance of the methods.

1. INTRODUCTION

In environmental epidemiology, the existence of exposure thresholds to some toxic reproductive and developmental agents is sometimes assumed by researchers ([2], [3], [8], [11]). More specifically, [11] pointed out that a threshold is a dose below which an outcome seen in excess of that is not produced, which can invalidate the non-threshold models. Therefore, instead of assuming a linear association between exposure and a health outcome, researchers sometimes assume a step-like relationship between them: the outcome takes on one distribution when the exposure is less than a certain dose, and takes on another when the exposure exceeds that. Usually the latter indicates a tighter exposure-disease association.

For a long time, people have been developing appropriate methods to find an exact threshold dosage. However, instead of making contributions to the topic of estimating the threshold dosage itself, this paper aims to take a different view of the threshold problem and propose methods for solving misclassification problems in threshold settings. In this regard, suppose exposures to a toxic agent are measured repeatedly over time, and the research question of interest involves the relationship between an adverse health outcome and an individual's true mean exposure over time. This is a reasonable hypothesis in cases where chronic exposure is deemed to make people unhealthy, rather than an acute one-time exposure. However, since the true mean exposure for a specific subject is unknown to researchers in most studies, a natural surrogate for it is the arithmetic average of the exposures observed. Therefore, measurement error or misclassification would occur when taking this surrogate to define the explanatory variable instead of the true mean exposure itself. The analytical problem becomes how to adjust for the misclassification and correct the exposure-outcome relationship with a proper model.

The motivating example for this paper is provided by a reproductive health study known as the Mount Sinai Study of Women Office Workers (MSSWOW). The prospective cohort study was conducted between 1991 and 1994 with our primary interest being the association between spontaneous abortion and repeated menstrual cycle length data derived from the women's diaries. However, when the subject-specific mean and variance describing each woman's cycle history are 'true' predictors of interest in models for spontaneous abortion, the sample mean and variance researchers usually utilize as surrogates will introduce measurement errors to the study. A previous paper [9] conducted careful analysis of this data and suggested that compared with 30- to 31-day cycles, women with shorter and longer cycles were more likely to experience spontaneous abortion. However, this paper did not address the misclassification issue as sample means and variances were "plugged in" to the health outcome models in the analysis. Such misclassification errors may not be ignorable in some cases, especially for those women with small numbers of cycles observed in the study. In particular, the use of surrogate sample means and variances as a basis for classifying women into "high" or "low" groups with regard to the theoretically true subject-specific mean cycle length and variability can introduce misclassification bias, when the research question is to assess the association between this group membership status and some health outcome.

This paper is organized as follows: Model assumptions and details of parameter estimation for categorical health outcomes will be presented in section 2. Simulation results are shown in section 3, followed by the motivating example analysis results in section 4 and a brief discussion.

*Biostatistics Dept., Emory University, Atlanta, GA

2. METHODS

2.1 Homogeneous Within-Subject Variances

2.1.1 Matrix Method

Suppose we are interested in the association between whether an individual's true mean exposure to a certain toxic agent exceeds a certain threshold, and a binary adverse health outcome. An example of such an outcome in the MSSWOW study is spontaneous abortion and the exposure of interest is the mean and/or variability of menstrual cycle length. Then since the mean exposure of a subject is unknown, we impose the generally reasonable assumption that a single measurement of exposure for a single subject, that the measurement fluctuates around that individual's mean exposure μ_i via a random normal disturbance ϵ_{ij}. Also, assume the mean exposure (μ_i) for subject i varies around the overall mean from the population μ via another random normal disturbance b_i, i.e.,

$$x_{ij} = \mu_i + \epsilon_{ij} \qquad i = 1, \ldots, k; \;\; j = 1, \ldots, n_i \tag{1}$$

and

$$\mu_i = \mu + b_i \quad i = 1, \ldots, k \tag{2}$$

where we assume $b_i \overset{iid}{\sim} N(0, \sigma_b^2)$, $\epsilon_{ij} \overset{iid}{\sim} N(0, \sigma_\epsilon^2)$, and b_i and ϵ_{ij} are mutually independent of each other. In this section, we assume homogeneous within-subject variance σ_ϵ; this assumption will be relaxed in section 2.2.

To simplify the model, we combine (1) and (2) into a single familiar linear mixed model as follows:

$$x_{ij} = \mu + b_i + \epsilon_{ij} \qquad i = 1, \ldots, k; \;\; j = 1, \ldots, n_i \tag{3}$$

To model the relationship between the adverse health outcome and whether the mean exposure exceeds the threshold, we apply a simple logistic regression model, i.e.,

$$\text{logit}[\text{P}(Y_i = 1)] = \delta_0 + \delta_1 I(\mu_i > t) \tag{4}$$

where t is the exposure mean threshold level, and $I(.)$ is an indicator function, which takes value one if the criteria in the parentheses is satisfied and zero if not. Since the true individual mean exposure μ_i is not available directly, a natural surrogate of that, denoted as $\tilde{\mu_i}$, would be the average of the repeated exposure measurements taken on subject i, i.e., $\tilde{\mu_i} = \frac{1}{n_i}\sum_{j=1}^{n_i} x_{ij}$. Hence, a 'naive' approach would take $I(\tilde{\mu_i} > t)$ as the explanatory variable in (4) instead of $I(\mu_i > t)$.

A special case in this setting is when we have balanced data so that every subject has the same number of exposure measurements, i.e., $n_1 = n_2 = \ldots = n_k = n$. In this restrictive but instructive case, we develop an approach as a special application of the matrix method [1] to correct for misclassification in binary exposure data.

In the current setting, the events 'truely exposed' and 'exposed prone to misclassification' (E=1 and W=1 in [1]) are correspond to '$\mu_i > t$' and '$\tilde{\mu_i} > t$', respectively. Although in our case no validation study are assumed available, we can still estimate the fundamental the sensitivity and specificity parameters based on the normal distribution assumptions in (1) and (2), as suggested previously [5]. More specifically,

$$sen = [\sqrt{2\pi}\sigma_b p]^{-1} \int_{t-\mu}^{\infty} \Phi[\sqrt{n}(b_i - (t-\mu))/\sigma_\epsilon] \text{exp}[-{b_i}^2/(2\sigma_b^2)] db_i \tag{5}$$

$$sp = [\sqrt{2\pi}\sigma_b (1-p)]^{-1} \int_{-\infty}^{t-\mu} \Phi[\sqrt{n}((t-\mu) - b_i)/\sigma_\epsilon] \text{exp}[-b_i^2/(2\sigma_b^2)] db_i \tag{6}$$

where $\bar{\epsilon}_i = \frac{1}{n}\sum_{j=1}^{n} \epsilon_{ij}$, $p = \Phi[(\mu - t)/\sigma_b]$.

If we insert estimates of *sen* and *sp* into the classical matrix method formular, and replace the conditional probability of misclassified exposure given disease status, i.e., $\text{P}(W = 1|D = 1)$, $\text{P}(W = 0|D = 1)$, $\text{P}(W = 1|D = 0)$, $\text{P}(W = 0|D = 0)$ in [1] with the corresponding observed count proportions, (e.g., $\hat{P}(W = 1|D = 1) = \hat{P}(\tilde{\mu_i} > t|D = 1)$), we obtain the estimated probabilities of true exposure given disease status $\text{P}(E = 1|D = 1)$, $\text{P}(E = 0|D = 1)$, $\text{P}(E = 1|D = 0)$, $\text{P}(E = 0|D = 0)$. The true odds ratio for the unobserved cells takes representation

$$\frac{\text{P}(\mu_i > t, y_i = 1) \times \text{P}(\mu_i < t, y_i = 0)}{\text{P}(\mu_i < t, y_i = 1) \times \text{P}(\mu_i > t, y_i = 0)} = \frac{\text{P}(E = 1|D = 1) \times \text{P}(E = 0|D = 0)}{\text{P}(E = 0|D = 1) \times \text{P}(E = 1|D = 0)}$$

Thus an adaptation of the matrix method provides a viable approach to the mean threshold problem in the special case where exposure data are balanced, the outcome is binary, and there are no covariates.

2.1.2 Likelihood Method

Given that the matrix method could only be adapted to estimate the odds ratio for balanced data, we next seek to derive a maximum likelihood method to handle this misclassification problem with unbalanced data.

With the TDM specified in (4), and the MEM dictated in (3), along with the non-differential measurement error assumption, we are able to write down the likelihood as

$$\begin{aligned}
& \mathcal{L}(\theta; \mathbf{Y}, \mathbf{X}) \\
= & \prod_{i=1}^{k} \Big(\int_{-\infty}^{\infty} [f(y_i|(b_i, x_{ij})) \times (\prod_{j=1}^{n_i} f(x_{ij}|b_i)) \times f(b_i)] db_i \Big) \\
= & \prod_{i=1}^{k} \Big(\int_{-\infty}^{\infty} [f(y_i|b_i) \times (\prod_{j=1}^{n_i} f(x_{ij}|b_i)) \times f(b_i)] db_i \Big) \qquad (7) \\
= & \prod_{i=1}^{k} \Big(\int_{t-\mu}^{\infty} [(\prod_{j=1}^{n_i} \frac{exp(-\frac{(x_{ij}-\mu-b_i)^2}{2\sigma_\epsilon^2})}{\sqrt{2\pi}\sigma_\epsilon}) \times [\frac{exp(\delta_0+\delta_1)}{1+exp(\delta_0+\delta_1)}]^{y_i} [\frac{1}{1+exp(\delta_0+\delta_1)}]^{(1-y_i)} \\
& \times \frac{exp(-\frac{b_i^2}{2\sigma_b^2})}{\sqrt{2\pi}\sigma_b}] db_i \\
+ & \int_{-\infty}^{t-\mu} [(\prod_{j=1}^{n_i} \frac{exp(-\frac{(x_{ij}-\mu-b_i)^2}{2\sigma_\epsilon^2})}{\sqrt{2\pi}\sigma_\epsilon}) \times [\frac{exp(\delta_0)}{1+exp(\delta_0)}]^{y_i} [\frac{1}{1+exp(\delta_0)}]^{(1-y_i)} \times \frac{exp(-\frac{b_i^2}{2\sigma_b^2})}{\sqrt{2\pi}\sigma_b}] db_i \Big)
\end{aligned}$$

Note that (7) holds under the non-differential measurement error, i.e., assuming that the observed measurements give no more information about the outcome once the true mean exposure is known. This follows from the definition of the TDM in (4). We are able to estimate $\theta = (\mu, \sigma_b, \sigma_\epsilon, \delta_0, \delta_1)$ by maximizing the integrated likelihood above using a Newton-Raphson algorithm. Available optimization routines such as the SAS NLPQN function in PROC IML can be used to conduct this optimization. The SAS procedure NLMIXED also allows user specification of the observation-specific likelihoods and optimizes the full likelihood via various optimization algorithms ([6], [7]). We conducted the analysis in both PROC IML and PROC NLMIXED to assess numerical reliability.

For balanced data, the matrix method with MLEs of exposure model parameters inserted into (5) and (6) is equivalent to the likelihood method but computationally less demanding. However, for unbalanced data, the matrix method is no longer available because the sensitivity and specificity calculated in (5) and (6) vary when n_i varies. Hence a constant expectation across different subjects is not available, which is required for the matrix method. On the other hand, the likelihood method is not subject to the restriction that the number of exposure measurements has to be equal across the subjects and it can also handle other covariates in the TDM and/or MEM. In other words, the likelihood method can be more generally applied than the matrix method. Another difference between the two methods is that the the matrix method can only provides an estimated crude odds ratio, i.e., addressing the crude association between the outcome and whether the individual mean exposure exceeds the threshold. For other parameters, such as possible covariate effects in the TDM, only the likelihood method can provide appropriate estimates.

2.2 Heterogeneous Within-Subject Variances

2.2.1 Likelihood Method

As our motivating example from the MSSWOW study requires modeling the variance of the menstrual cycle length, in this section we generalize the homogeneous within-subject variance assumption proposed in (3) in section 2.1. Instead of assuming a constant within-subject variance σ_ϵ for the error term ϵ, we assume the within-subject variance v_i follows a lognormal distribution, as proposed in prior research [4], i.e.,

$$X_{ij} = \mu + b_i + \epsilon_{ij} \qquad i = 1, \ldots, k; j = 1, \ldots, n_i$$

where

$$\epsilon_{ij}|v_i \overset{iid}{\sim} N(0, v_i), \qquad log(v_i) \overset{iid}{\sim} N(\alpha, \phi^2) \qquad (8)$$

Therefore, the log-normally distributed v_i accounts for randomness in variability of women's cycle lengths (X_{ij}) in the same way that b_i accounts for randomness in the mean. The advantage of this extension is that it allows the true within-subject variances in cycle length (v_i), as well as the mean cycle lengths ($\mu_i = \mu + b_i$), to vary randomly

across subjects. In the likelihood representation, v_i can be integrated out in the same way that b_i is in (8). For example, if we look at the exposure data only, expressing the likelihood of observing the repeated exposures x_{ij} for each subjects requires a double integration as in the following:

$$\mathscr{L}(\theta; \mathbf{X}) = \prod_{i=1}^{k} \{ \int_0^\infty \int_{-\infty}^\infty f(x_{ij}|b_i, v_i) f(b_i|v_i) f(v_i) db_i dv_i \}$$

To allow for possible correlation between b_i and v_i are correlated, we presume b_i and $log(v_i)$ to be bivariate normally distributed [4], i.e.,

$$\begin{bmatrix} b_i \\ log(v_i) \end{bmatrix} \sim N \left[\begin{pmatrix} 0 \\ \alpha \end{pmatrix}, \begin{pmatrix} \sigma_b^2 & \sigma_{bv} \\ \sigma_{bv} & \phi^2 \end{pmatrix} \right] \tag{9}$$

where here σ_{bv} is the covariance between b_i and $log(v_i)$.

In order to incorporate both subject-specific mean and variance in a disease-exposure relationship, we consider generalized linear models of the following type:

$$g[E(Y_i)] = \delta_0 + \delta_1 I(\mu_i > \mu) + \delta_2 I(v_i > e^\alpha) + \delta_3 I(\mu_i > \mu) I(v_i > e^\alpha)$$

where the threshold for mean cycle length is set at the overall mean μ, and the threshold for cycle length variance is set at its population median, e^α. Here, g is the link function accommodating different types of outcomes (e.g., continuous, binary, count etc.)

The structure of the threshold indicator could vary according to different research questions. For example, in the MSSWOW study, where the exposure pertains to women's menstrual cycles, researchers believe the risk of developing an adverse reproductive outcome is associated with abnormally long or short menstrual cycle length, and/or anomalously large variation in the cycle length [9]. To be consistent with this hypothesis, the above generalized linear model can be adjusted to take the form:

$$g(E(y_i)) = \delta_0 + \delta_1 I(t_1 < \mu_i < t_2) + \delta_2 I(v_i < \omega) + \delta_3 I(t_1 < \mu_i < t_2) I(v_i < \omega) \tag{10}$$

where t_1, t_2 are the lower and upper bound of the normal lengths of menstrual cycles, and ω is the threshold for the normal variation in the lengths. All of those may be set as fixed known constants prior to the analysis, e.g., based on prior research [9].

Although previous investigaors [9] conducted an epidemiologically sophisticated study that involved careful statistical modelling, from a measurement error/misclassification perspective their approach is consistent with what might be termed a 'naive' method. That is, the models used were analogous to model (10), except with the sample mean ($\bar{x}_i$) and sample variance (s_i^2) of cycle lengths replacing the true quantities (μ_i and v_i).

The likelihood considering both disease outcome and the exposure measurements will be as follows:

$$\mathscr{L}(\mathbf{\Theta}; \mathbf{Y}, \mathbf{X}) = \prod_{i=1}^{k} \{ \int_0^\infty [\int_{-\infty}^\infty f(y_i|x_{ij}, b_i, v_i; \theta) f(x_{ij}|b_i, v_i; \theta) f(b_i|v_i; \theta) f(v_i; \theta) db_i] dv_i \}$$

where $\mathbf{\Theta}$ here includes all the parameters to be estimated, i.e., $(\delta_0, \delta_1, \delta_3, \sigma_e, \mu, \sigma_b, \alpha, \phi)$.

As with our previous work to adjust for misclassification, the advantage of this likelihood representation is that it allows direct estimates of and inference about the association between true cycle length means and variances and the outcome of interest.

2.2.2 Empirical Bayes (EB) Method

As an alternative to the naive method, where the subjects are grouped based on the raw sample quantities ($\bar{x}_i$ and s_i^2), the EB method will utilize the empirical Bayes predictors for μ_i and $log(v_i)$ as surrogates for the true classifications. These EB predictions are estimates of the corresponding posterior means, i.e.,

$$\tilde{\mu}_{i,B} = E(\mu_i|X_i) \quad and \quad \tilde{\eta}_{i,B} = E[ln(v_i)|X_i]$$

where $\eta_i = ln(v_i)$, and subscript B indicates the EB estimates. We obtain the EB predictions by utilizing the standard software, e.g., NLMIXED in SAS ([6], [7]).

Prior work [4] pointed out the tremendous amount of shrinkage by the EB estimates in comparison with the sample quantities. Therefore, sensitivity would generally be lower but specificity would be higher when basing mean or variance classification relative to threshold on the EB predictions as the surrogates, rather than relying on the sample quantities. The details on comparing the two surrogates will be discussed in the simulation section.

3. SIMULATION RESULTS

3.1 Homogeneous Within-Subject Variances

3.1.1 Matrix Method

To demonstrate the performance of the matrix method adaptation, we simulated 500 datasets, each with a sample size of 300. The true parameter values are $(\delta_0, \delta_1, \mu, \sigma_b, \sigma_\epsilon) = (1, 1.5, 2.5, 1, 1.5)$ with the threshold t as 2.5. The number of repeated exposure measurements is 5 for each subject in the simulation.

Since we are interested in the association between the health outcome and whether the true mean exposure (μ_i) exceeds the threshold, and the matrix method can only provide direct estimates of the odds ratio, the simulation results in Table 1 will present estimates only for δ_1 only based on 500 simulations. Also in Table 1, five situations of different combinations of the overall mean and between- and within-subject variance components are listed to show how those parameters affect the sensitivity and specificity and the limiting value (δ_1^*) of the naive estimators. The threshold t remains constant at 2.5 throughout all situations.

Table 1: Simulation results and hypothetical situations illustrating the matrix method for binary outcome: simulation results based on 500 simulated datasets, sample size = 300, t=2.5, number of replicates = 5, and $(\delta_0, \delta_1) = (0, 1)$

μ	σ_b	σ_ϵ	SEN	SPC	δ_1^*	$\delta_1(SD)$
2.5	1	1.5	0.81	0.81	0.61	1.01 (0.41)
2.5	1	2.5	0.74	0.74	0.47	1.02 (0.53)
2.5	2	1.5	0.90	0.90	0.79	1.01 (0.32)
1.5	1	1.5	0.75	0.90	0.48	1.03 (0.75)
3.5	1	1.5	0.90	0.75	0.54	0.99 (0.57)

From Table 1 we can see that the matrix method performs well, in the sense that the bias and the empirical standard deviation of the parameter δ_1 are reasonably small. Figures (not shown here) plotting sensitivity and specificity versus one of the parameters μ, σ_b and σ_ϵ at a time reveal the fact that sensitivity and specificity increase as the between-subject variance σ_b increases. The intuition behind this is that when the between-subject variance is large, true subject-specific mean exposures of each subject tend to be quite different from the threshold so that it would be easier to detect whether the mean exceeds the threshold or not. However, the sensitivity and specificity decrease when the within-subject variance increases. This is also reasonable because when the within-subject variance is relatively larger, the true mean exposure becomes harder to approximate by the surrogate ($\bar{x_i}$) which makes the sensitivity and specificity relatively lower. These trends could also be perceived in formulas (5) and (6).

3.1.2 Likelihood Method

For demonstrative and comparison purposes, we also assume balanced data in this section with the same simulation settings applied for the matrix method. The simulation results are shown in Table 2.

Table 2: Simulation illustrating the likelihood approach for binary outcome: results based on 500 simulated datasets, sample size = 300, t=2.5, number of replicates = 5, and $(\delta_0, \delta_1, \mu, \sigma_b, \sigma_\epsilon) = (0, 1, 2.5, 1, 1.5)$

Parameters	True values	Mean Est.	Emp.SD	Mean Est. SE	95% coverage
δ_0	0.00	0.000	0.23	0.20	0.92
δ_1	1.00	1.000	0.38	0.36	0.95
μ	2.50	2.505	0.07	0.07	0.94
σ_b	1.00	1.001	0.06	0.05	0.96
σ_ϵ	1.50	1.506	0.03	0.03	0.94

As we can see from Table 2, the bias associated with each MLE is quite small, indicating that the estimates are close to the true parameter values on average. The empirical standard deviations are not far off from the means of the estimated standard errors. The 95% coverages of the parameters δ_0, δ_1 and μ are close to 95%.

3.2 Heterogeneous Within-Subject Variances

Analogous to the real data example situation (details will be presented in the next section), and based on the model described in section 2.2, we conducted simulation where the outcome is binary, the variances of the exposure for each subject are considered to be different, and the mean and the variance of the exposures are assumed to be correlated. Simulation results are shown in Table 3

Table 3: Simulation illustrating the likelihood approach for modeling correlated mean and variance: results based on 500 simulated datasets, sample size = 800, threshold is [26.90, 30.65], number of replicates = 8, and $(\delta_0, \delta_1, \delta_2, \mu, \sigma_b, \alpha, \phi, \sigma_{bv})$ = (-0.8, -0.4, 0.3, 28, 2.8, 1.8, 1.0, 2.0)

	δ_0	δ_1	δ_2	μ	σ_b	α	ϕ	σ_{bv}
Results Based on Likelihood Method								
True Values	-0.8	-0.4	0.3	28	2.8	1.8	1.0	2.0
Mean Est.	-0.812	-0.371	0.302	27.982	2.742	1.803	0.970	1.902
Emp.SD	0.191	0.230	0.230	0.180	0.140	0.060	0.034	0.211
Mean Est. SE	0.190	0.215	0.221	0.093	0.053	0.038	0.031	0.151
95% coverage	95.4%	93.9%	95.8%	–	–	–	–	–
Results Based on EB Method								
True Values	-0.8	-0.4	0.3	–	–	–	–	–
Mean Est.	-0.801	-0.298	0.208	–	–	–	–	–
Emp.SD	0.147	0.167	0.161	–	–	–	–	–
Mean Est. SE	0.151	0.157	0.160	–	–	–	–	–
95% coverage	96.6%	89.0%	90.6%					
Results Based on Naive Method								
True Values	-0.8	-0.4	0.3	–	–	–	–	–
Mean Est.	-0.814	-0.286	0.140	–	–	–	–	–
Emp.SD	0.274	0.170	0.279	–	–	–	–	–
Mean Est. SE	0.263	0.156	0.270	–	–	–	–	–
95% coverage	95.6%	89.3%	84.4%	–	–	–	–	–

The simulation results show that the estimates of the parameters from the likelihood are reasonably close to the true parameter values, and the empirical standard deviations are close to the mean estimated standard errors. The 95% Wald confidence interval coverages for the model coefficients are near nominal, all of which suggest the favorable performance of the likelihood method in modeling both correlated mean and variance with the heterogeneous subject-specific variation assumption. Compared to the likelihood estimates, those from both the EB method and the naive method tend to attenuate the effect to the null, which is consistent with the findings from much of the measurement error literature. We observe from the simulation that the method of "plugging in" the EB predictors still introduces misclassification and does not guarantee smaller bias than the naive method, which makes sense given that sensitivity is generally decreased and specificity increased via this approach. Although prior work [10] found for continuous exposures that the use of the surrogate of EB predictors performs similarly to the likelihood method, our simulation confirms the expected result that the estimates from dichotomizing the independent variable based on the EB predictors does not provide a consistent estimator.

4. REAL-DATA EXAMPLE

To demonstrate the proposed method, we get back to the motivating example of the MSSWOW study introduced in Section 1. The research objective here is to examine whether women with menstrual cycle lengths within a normal range, and/or with cycle length variation less than a certain threshold, are at lower risk of experiencing spontaneous abortion. Among a total number of 470 with menstrual cycle length information in the study (see [9] for details), 162 women are included in this analysis with pregnancy outcome of either live birth (118 women [73%]) or spontaneous abortion (44 women [27%]), including subclinical spontaneous abortion, blighted ovum and clinical spontaneous abortion. For those women with more than one pregnancy outcome, only the first outcome and cycles up to the first outcome are included in the analysis. We begin with a preliminary look into the exposure of menstrual cycle data with heterogeneous within-subject variance model structure when the mean and variance are correlated, as in (3), (8) and (9).

Using the estimated nuisance parameters from the exposure model (results not shown), we dichotomize the subjects' mean and variation of menstrual cycle lengths based on whether their mean cycle length in days is within the normal range of [26.90, 30.65] (corresponding to the 25th and 75th percentile of the normal distribution with mean and variance from the estimated parameters) and whether the variance exceeds the threshold of 4.8 (mean of the lognormal distribution from the estimated parameters). The threshold cut-offs for the mean are different from [9], due to taking advantage of the estimates from the exposure model with random effects. We apply the same thresholds across the three methods (likelihood method, EB method and naive method) for comparison purposes, although in the absence of a random effects exposure model, people often utilize sample quantiles as the cut-off points.

Regarding the outcome model, the interaction between the mean and variance is perceived to be insignificant in our data example, and herefore is left out of the model. In addition to the dichotomized mean and variance of the cycle lengths as the independent variable, we control in the model for dichotomized maternal age by the cutoff of 35 years of age. In the analysis, we built models with cycle mean only (results not shown), cycle variance only (results not shown) and both mean and variance in the model. Table 4 presents the estimates for those models from the likelihood method, empirical Bayes method and the naive method.

Table 4: Estimated odds ratio for the association between spontaneous abortion and whether the menstrual cycle mean and variance exceed their respective thresholds based on MSSWOW Data

	Results based on likelihood model		Results based on Emp. Bayes method		Results based on naive model	
Risk Factors	OR	95% Wald CI	OR	95%Wald CI	OR	95% Wald CI
Mean of the cyles ([29,33] vs. (<29 or >33))	0.22	0.05, 0.97	0.42	0.20, 0.92	0.41	0.17, 0.99
Variance of the cycles (<23 vs. ≥23)	0.69	0.20, 2.39	0.68	0.31, 1.50	0.57	0.26, 1.24
Maternal Age (<35 vs. ≥35)	0.56	0.22, 1.47	0.61	0.25, 1.47	0.62	0.25, 1.49

The analysis results show that those women with cycle mean within the middle two quantitles [26.90, 30.65] bear lower probability of experiencing spontaneous abortion than those with cycle mean shorter or longer than the normal range, suggested consistently by the likelihood method, EB method and naive method (see Table 4 for partial results). In terms of the magnitude of the estimated effects, the naive method and the EB method both appear to attenuate the effect, compared to the likelihood method. On the other hand, analyses also reveal that the dichotomized cycle variability appears to be an insignificant risk factor for spontaneous abortion as implied by all methods. The estimated variance effect from the models with only dichotomized cycle variance and maternal age is slightly further from the null by the naive method as compared to the likelihood method, while the EB has the estimated effect closest to the null. However, with mean cycle length and maternal age controlled for in the model, the magnitude of the estimated variance effects are all similar across the three methods (Table 4). The same occurs for maternal age, i.e., it is an insignificant covariate with very similar estimated effects from all the methods after adjusting for the cycle means and variability. The confidence intervals stemming from the EB method are narrower than those from the likelihood method, likely due to the fact that the variances of the EB predictions were not incorporated into inferences based on the second stage model, where the predictions are "plugged in" as surrogates.

References

[1] B.A. Barron. The effects of misclassification on the estimation of relative risk. *Biometrics*, 33:414–418, 1977.

[2] J.K. Haseman and L.L. Kupper. Analysis of dichotomous response data from certain toxicological experiments. *Biometrics*, 35:281–292, 1979.

[3] T.F. Hatch. Thresholds: Do they exist? *Archives of Environmental Health*, 22:687–689, 1971.

[4] R.H. Lyles, A. Munõz, J. Xu, J.M.G. Taylor, and S.J. Chmiel. Adjusting for measurement error to assess health effects of variability in biomarkers. *Statistics in Medicine*, 18:1069–1086, 1999.

[5] R.H. Lyles and J. Xu. Classifying individuals based on predictors of random effects. *Statistics in Medicine*, 18:35–52, 1999.

[6] SAS Institute Inc. *SAS/IML Software: Changes and Enhancements, Release 8.2*. SAS Institute Inc, Cary, NC, 2001.

[7] SAS Institute Inc. *SAS/Share 9.2 User's Guide*. SAS Institute Inc, Cary, NC, 2008.

[8] P.F. Schwartz, C. Gennings, and V.M. Chinchilli. Threshold models for combination data from reproductive and developmental experiments. *Journal of the Ametican Statistical Association*, 90(431):862–870, Sep. 1995.

[9] C.M. Small, A.K. Manatunga, M. Klein, H.S. Feigelson, C.E. Dominguez, R. McChesney, and M. Marcus. Menstrual cycle characteristics - associations with fertility and spontaneous abortion. *Epidemiology*, 17(1):52–60, Jan 2006.

[10] K Wannemuehler. *Likelihood-based Measurement Error Adjustments in Occupational and Environmental Exposure Studies*. PhD thesis, Biostatistics Dept., Emory University, 2007.

[11] J.G. Wilson. *Environmental and Birth Defects*. New York: Academic Press, 1973.