Published in final edited form as:

Stat Med. 2016 April 30; 35(9): 1502–1513. doi:10.1002/sim.6798.

Estimation of interaction effects using pooled biospecimens in a case-control study

Michelle D. Danaher^{a,b}, Paul S. Albert^a, Aninyda Roy^b, and Enrique F. Schisterman^{a,*}

^aEunice Kennedy Shriver National Institute of Child Health and Human Development, Bethesda, Maryland, U.S.A

^bDepartment of Mathematics and Statistics, University of Maryland, Baltimore County, Baltimore, Maryland, U.S.A

Abstract

Pooling, or physically mixing biospecimens, prior to evaluating biomarkers dramatically reduces biomarker evaluation cost, reduces the quantity of biospecimens required of each individual, and may reduce the percentage of laboratory measurements below the lower limit of detection. Motivated by a case control study on miscarriage (binary outcome) and cytokines (continuous exposures), we are interested in estimating parameters in a logistic regression, where individuals with the same disease status (with or without a miscarriage) are paired and their pooled cytokine concentrations are assessed. Previous research has proposed a set-based logistic model to evaluate the relationship between a disease and pooled exposures. While the set-based logistic model is very useful for estimating main effects, it cannot estimate interactions of continuous exposures when both are measured in pools. Therefore, we propose using the Expectation Maximization (EM) algorithm to obtain estimators of all parameters in logistic regression model, including interactions effects. Using a simulation study, we present comparisons of efficiency under different scenarios where exposures have been measured in pools and individually. Our simulations show that randomly sampling half of the available biospecimens has less efficiency than pooling pairs of biospecimens stratified by disease status. The EM algorithm provides a method for estimating interaction effects when biospecimens have already been pooled for other reasons such as the gain in efficiency for estimating main effects demonstrated by previous research. This manuscript demonstrates that the EM algorithm offers a promising approach to estimate interaction effects of pooled biospecimens. Copyright

Keywords

cytokines; expectation maximization; logistic regression; pooling designs; skewed biomarkers

1. Introduction

Expensive biomarkers often limit investigators' ability to test interesting hypotheses. For example, measuring polychlorinated biphenyls (PCBs) costs between \$500 and \$1000 per

*Correspondence to: 6100 Executive BLVD Room 7B03H, MSC 7510, Bethesda Md 20892-7510. schistee@mail.nih.gov.

sample [1]. Due to the prohibitive cost of biomarker evaluation, investigators may choose to measure the expensive biomarker in a small sub-sample from their population of interest. However, this approach is unsatisfactory due to the lost information from the reduction in sample size. In an attempt to reap more information within financial constraints, a number of investigators have proposed measuring biomarkers using biospecimen pools, rather than measuring each individual's biospecimen [2, 3, 4, 5]. Biospecimen pools are physical mixtures of equal volumes of two or more individuals' biospecimens. Measuring pooled biospecimens with two individuals each, rather than measuring each individual's biospecimen, reduces the cost associated with biomarker assessment up to fifty percent. Another benefit is that pooling may reduce the quantity of biospecimen required from each individual, thus saving invaluable biological resources [6]. Furthermore, when measuring a continuous biomarker, pooling may reduce the percentage of measurements under the limit of detection [2, 4, 5, 7], and may help to preserve an individual's privacy.

Our interest in biospecimen pooling is motivated by an epidemiological study on miscarriage and chemokines. Chemokines, which are involved in the regulation of inflammation and immune response, are hypothesized to be related to unexplained pregnancy losses [8, 9]. To explore this hypothesis, samples were obtained from the Collaborative Perinatal Project (CPP) cohort, which was conducted from 1959 to 1965. Biospecimens were assessed in 358 women who had a miscarriage and 378 controls matched by gestational age at sample collection. Details regarding sample collection and demographics are provided elsewhere [10, 9, 11]. An important feature of this biomarker study is that individual measurements of the chemokines macrophage inhibitory protein MIP-1 α and MIP-1 β were obtained, as well as pooled measurements of two individuals' biospecimens matched on case status. Therefore, we have the unique ability to make comparisons between pooled measurements and the traditional individual measurements.

In the context of logistic regression with pooled exposure assessments matched on case status, Weinberg and Umbach [4, 5] proposed a set-based logistic model, which is a very useful tool for estimating main effects in a logistic regression model. However, when an investigator is interested in exploring a relationship between a disease (i.e. miscarriages) and exposures (i.e. chemokines) it is likely that they will want to explore possible interactions and/or higher-order effects (e.g. the interaction between two chemokines). The set-based logistic model is not useful for these higher-order effects because the pooled measurement is assumed to be the average of the individual measurements, and the product of sums is not the same as the sum of products. A regression calibration approach, as proposed by Zhang and Albert [6] may be adapted to handle interaction terms when both exposures are measured in pools. However this approach requires identifying a good predictor for regression calibration, and therefore is unsatisfactory when additional covariates are not available to assist in predicting individual level exposure. If an investigator measures exposures in pools to gain efficiency in estimating the main effects as demonstrated by Weinberg and Umbach [4, 5], currently there is no recommended method for them to explore possible interaction effects using their pooled data. In this paper we propose using the Expectation Maximization (EM) [12] algorithm to estimate interaction effects in a logistic model by treating a fraction of the exposure data as missing data. In particular, as in

our motivating epidemiological study, we will explore the situation where two individuals' biospecimen are mixed in each pool, therefore half of the data is missing.

In Section 2 we define some notation and distributional assumptions used throughout the paper. In Section 3 we provide the likelihood, the EM algorithm, and an estimator for the variance of the parameters estimated using the EM algorithm. In Section 4 we use a simulation study to evaluate the performance of the pooling design compared with the traditional design where everyone has individual measurements of exposure. In Section 5 we illustrate our method by estimating the interaction between MIP-1 α and MIP-1 β on miscarriage using pooled exposure assessments and compare those results to the estimated interaction using individual exposure assessments. Lastly, we conclude with a discussion in Section 6.

2. Notation and assumptions

Let n be the number of individuals in the study, n^p be the number of pools in the study, and for illustration purposes, we assume that there are only two individuals' biospecimens in each pool (this is consistent with our motivating example, though it is straightforward to extend the methodology to larger pool sizes). There are $n^p = n/2$ pools of biospecimens, where we assume that n^p is an integer. Let $\mathbf{x}_i^p = \mathbf{x}_{i1} + \mathbf{x}_{i2}$, where $\mathbf{x}_{ij} = [x_{ij1}, x_{ij2}]'$ are the exposures (e.g. concentrations of MIP-1 α and MIP-1 β in the motivating example) for the j^{th} person in the i^{th} pool. The observed exposure within a pool is the average of the individuals' exposures, $\mathbf{x}_i^p/2$; therefore $\mathbf{x}_i^p$ is the pooled observation times two. Let the binary response for the j^{th} person in the i^{th} pool be y_{ij} , which is equal to 1 if the individual has the disease (e.g. miscarriages), and 0 if the individual does not have the disease (control). We assume that the disease status given the exposures has the following distribution,

$$y_{ij} | \mathbf{x}_{ij}, \boldsymbol{\beta} \stackrel{\text{ind}}{\sim} \text{Bernoulli}[\pi(\mathbf{x}_{ij}, \boldsymbol{\beta})],$$

for $i = 1, 2, \dots, n^p, j = 1, 2$, where $\stackrel{\text{ind}}{\sim}$ means independently distributed, and $\pi(\mathbf{x}_{ij}, \boldsymbol{\beta})$ is the logistic link given by,

$$\pi(\mathbf{x}_{ij}, \boldsymbol{\beta}) = \frac{\exp[\boldsymbol{\beta}' D(\mathbf{x}_{ij})]}{1 + \exp[\boldsymbol{\beta}' D(\mathbf{x}_{ij})]},$$

$\boldsymbol{\beta} = [\beta_0, \beta_1, \beta_2, \beta_3]'$ are the unknown parameters of interest, $D(\mathbf{x}_{ij}) = [1, x_{ij1}, x_{ij2}, x_{ij1} \times x_{ij2}]'$. Additional covariates may be included in the logistic regression model, either measured in pools or measured individually. For illustrative purposes, we chose to only include the main effects and interaction term of the pooled exposures, however adding additional covariates is straightforward.

In logistic regression, a distributional assumption for the exposures, $\mathbf{x}_{ij}$, is not required. However, the EM algorithm requires a distributional assumption for the missing individual

exposure information (i.e. we observe the exposures measured in pools, rather than individual measurements). We would like to be flexible in choosing a distribution for the exposures, therefore, since biomarkers tend to be positive and highly skewed (which is seen in our example with MIP-1 α and MIP-1 β chemokines) we assume that there exists a Box-Cox transformation [13] for the exposures which approximates a normal distribution,

$$t(\mathbf{x}_{ij}, \boldsymbol{\lambda}) \overset{\text{iid}}{\sim} N_2(\boldsymbol{\mu}, \boldsymbol{\Sigma}),$$

where $\overset{\text{iid}}{\sim}$ means approximately independently identically distributed, $\boldsymbol{\lambda} = [\lambda_1, \lambda_2]'$ $\mathbf{0}$, $\mathbf{x}_{ij} > \mathbf{0}$, and t is the Box-Cox transformation of the exposures, $\mathbf{x}$,

$$t(\mathbf{x}_{ij}, \boldsymbol{\lambda}) = \begin{bmatrix} (x_{ij1}^{\lambda_1} - 1)/\lambda_1 \\ (x_{ij2}^{\lambda_2} - 1)/\lambda_2 \end{bmatrix}.$$

Define $J_{\mathbf{x}}$ to be the Jacobian for this transformation,

$$J_{\mathbf{x}}(\mathbf{x}_{ij}, \boldsymbol{\lambda}) = x_{ij1}^{\lambda_1 - 1} x_{ij2}^{\lambda_2 - 1}.$$

Therefore, $f_{\mathbf{x}}$ is the marginal distribution of the exposures,

$$f_{\mathbf{x}}(\mathbf{x}_{ij} | \boldsymbol{\lambda}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = J_{\mathbf{x}}(\mathbf{x}_{ij}, \boldsymbol{\lambda}) \phi(t(\mathbf{x}_{ij}, \boldsymbol{\lambda}) | \boldsymbol{\mu}, \boldsymbol{\Sigma}) I(\mathbf{x}_{ij} > \mathbf{0}),$$

where I is an indicator function, ϕ is the bivariate normal density,

$$\phi(t(\mathbf{x}_{ij}, \boldsymbol{\lambda}) | \boldsymbol{\mu}, \boldsymbol{\Sigma}) = (2\pi)^{-1} |\boldsymbol{\Sigma}|^{-1/2} \exp \left[-1/2 (t(\mathbf{x}_{ij}, \boldsymbol{\lambda}) - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (t(\mathbf{x}_{ij}, \boldsymbol{\lambda}) - \boldsymbol{\mu}) \right],$$

and $\boldsymbol{\Sigma}$ is parameterized as follows,

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & \sigma_1 \sigma_2 \rho \\ \sigma_1 \sigma_2 \rho & \sigma_2^2 \end{pmatrix}.$$

For notational convenience we define the parameters associated with the marginal distribution of the exposure by $\boldsymbol{\gamma} = [\boldsymbol{\lambda}', \sigma_1, \sigma_2, \rho, \boldsymbol{\mu}']'$, and we define the unknown parameters by $\boldsymbol{\theta} = [\boldsymbol{\beta}', \boldsymbol{\gamma}']'$.

It is important to note that the Box-Cox transformed normal distribution is used to find a flexible distribution for the exposures; we are not using the Box-Cox transformation to

transform the exposures in the logistic model, therefore interpretations of the logistic model parameters will not be affected. Furthermore, while we choose the Box-Cox transformed normal distribution for the exposures measured in pools, changing f_x is simple when there is reason to assume another distribution such as a multivariate gamma or normal distribution.

3. A case-control study

In a case-control study individuals are sampled from a cohort depending on their disease status, then exposure information is ascertained. Under the distributional assumptions defined in Section 2, the likelihood for a case-control study with complete information on every individual (i.e. no pooled measurements) is as follows,

$$L(\theta|\mathbf{y}_1, \mathbf{y}_2, X_1, X_2, \mathbf{s}_1, \mathbf{s}_2) = \prod_{i=1}^{n^p} \prod_{j=1}^2 Pr(Y=y_{ij}, \mathbf{x}_{ij}|s_{ij}=1, \theta),$$

where $\mathbf{y}_j = [y_{1j}, \dots, y_{npj}]'$, $X_j = [\mathbf{x}_{1j}, \dots, \mathbf{x}_{npj}]'$, $\mathbf{s}_j = [s_{1j}, \dots, s_{npj}]'$ for $j = 1, 2$, and s_{ij} is an indicator which equals one when the j^{th} person in the i^{th} pool is sampled. The sampling indicator depends only on the disease status, y_{ij} . Therefore the exposure, $\mathbf{x}_{ij}$, is conditionally independent of the sampling indicator, s_{ij} , given the disease status, y_{ij} . The case-control likelihood can be written as follows,

$$L(\theta|\mathbf{y}_1, \mathbf{y}_2, X_1, X_2, \mathbf{s}_1, \mathbf{s}_2) = \prod_{i=1}^{n^p} \prod_{j=1}^2 \frac{Pr(Y=y_{ij}|\mathbf{x}_{ij}, \boldsymbol{\beta}) f_x(\mathbf{x}_{ij}|\boldsymbol{\lambda}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) \tau(y_{ij})}{\alpha(y_{ij}, \boldsymbol{\theta})}, \quad (1)$$

where $\tau(y_{ij}) = Pr(Y = y_{ij}|s_{ij} = 1)$, is a design constant (e.g. in a case-control design with half cases and half controls, $\tau(y_{ij}) = 0.50$ for all i, j), and $\alpha(y_{ij}, \boldsymbol{\theta})$ is the probability of having the disease given the parameters, $\boldsymbol{\theta}$.

$$\alpha(y_{ij}, \boldsymbol{\theta}) = \int_{\mathbf{x} \in \mathbb{R}^{2+}} Pr(Y=y_{ij}|\mathbf{x}, \boldsymbol{\beta}) f_x(\mathbf{x}|\boldsymbol{\lambda}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) d\mathbf{x}.$$

For additional derivation clarification of the likelihood defined in Equation 1 see Supplementary Web Appendix A.

We assume that pooling is random within strata of the disease (i.e. $\mathbf{y}_1 = \mathbf{y}_2$). For notational convenience, define the function g as follows,

$$g(\mathbf{x}_{i1}, y_{i1}, \mathbf{x}_i^p, \boldsymbol{\theta}) = Pr(Y=y_{i1}|\mathbf{x}_{i1}, \boldsymbol{\beta}) f_x(\mathbf{x}_{i1}|\boldsymbol{\lambda}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) Pr(Y=y_{i1}|\mathbf{x}_i^p - \mathbf{x}_{i1}, \boldsymbol{\beta}) f_x(\mathbf{x}_i^p - \mathbf{x}_{i1}|\boldsymbol{\lambda}, \boldsymbol{\mu}, \boldsymbol{\Sigma}).$$

We can rewrite the likelihood defined in Equation 1 as,

$$L(\boldsymbol{\theta}|\mathbf{y}_1, X_1, X^p, \mathbf{s}_1) = \prod_{i=1}^{n^p} \frac{g(\mathbf{x}_{i1}, y_{i1}, \mathbf{x}_i^p, \boldsymbol{\theta}) [\tau(y_{i1})]^2}{[\alpha(y_{i1}, \boldsymbol{\theta})]^2}, \quad (2)$$

where $X^p = [\mathbf{x}_1^p, \dots, \mathbf{x}_{n^p}^p]'$, and $\mathbf{s}_1 = \mathbf{s}_2 = \mathbf{1}$ (i.e. if at least one biospecimen in a pool is a sampled, it implies that all biospecimens in the pool were sampled).

Direct maximization of the likelihood defined in Equation 2 (and equivalently Equation 1) is only possible when an estimate for the probability of having the disease ($\alpha(y_{ij}, \boldsymbol{\theta}) = \Pr(Y = y_{ij}|\boldsymbol{\theta})$) is obtainable. Define κ to be an estimate of the disease prevalence. An investigator may obtain an estimate of the disease prevalence, κ , from a cohort study, prior research, or expert opinion. Recall Prentice and Pyke [14] showed that the missing information on the disease prevalence only biases the intercept in a logistic regression model. However we are not able to directly maximize the likelihood assuming the traditional logistic link function because we have missing information on the exposures (due to the pooled sampling design); therefore the results provided by Prentice and Pyke [14] are not valid in this situation.

Once the disease prevalence is estimated, we can evaluate β_0 as a function of the disease prevalence estimate, κ , and the remaining parameters, $[\beta_1, \beta_2, \beta_3, \gamma']'$. To estimate β_0 we need to solve the following equation,

$$0 = \kappa - \alpha(1, \boldsymbol{\theta}), \quad (3)$$

for β_0 , where $\alpha(1, \boldsymbol{\theta})$ can be approximated using Gauss-Legendre quadratures (see Supplementary Web Appendix B). We approximate a solution to Equation 3 using the uniroot function in R [15]. Define β_0^* to be the function of κ and the remaining parameters $[\beta_1, \beta_2, \beta_3, \gamma']'$ which solves Equation 3, and define $\boldsymbol{\beta}^* = [\beta_0^*, \beta_1, \beta_2, \beta_3]'$ and $\boldsymbol{\theta}^* = [\boldsymbol{\beta}^{*'}, \gamma']$. All parameters are identifiable when we replace $\boldsymbol{\theta}$ for $\boldsymbol{\theta}^*$ in the likelihood defined in Equation 2.

3.1. The EM algorithm

When biospecimens are pooled, we do not observe complete data for the exposure, rather we observe the pooled exposure X^p , and if we knew X_1 we could recover $X_2 = X^p - X_1$. Due to missing information, direct maximization of the likelihood in Equation 2 is difficult, and therefore we propose using the EM algorithm.

The first step in the EM algorithm is to evaluate Q , which is the expected value of the log likelihood given our observed data, $\mathbf{y}_1$, X^p and $\mathbf{s}_1 = \mathbf{1}$, and an estimate of $\boldsymbol{\theta}^*$, say $\boldsymbol{\theta}^{*t}$,

$$Q(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{*t}) = E_{f_{x^m}} \{ \log [L(\boldsymbol{\theta}^*|\mathbf{y}_1, X_1, X^p, \mathbf{s}_1)] \} \\ = \sum_{i=1}^{n^p} 2\log[\tau(y_{i1})] - 2\log[\alpha(y_{i1}, \boldsymbol{\theta}^*)] + \int_0^{\mathbf{x}_i^p} \log[g(\mathbf{x}_{i1}, y_{i1}, \mathbf{x}_i^p, \boldsymbol{\theta}^*)] f_{x^m}(\mathbf{x}_{i1}|\mathbf{y}_1, X^p, \boldsymbol{\theta}^{*t}) d\mathbf{x}_{i1}$$

where f_{x^m} denotes the conditional distribution of the missing exposure data, given the disease status, the pooled exposures, and the current estimate of the parameters, θ^{*t} , that is,

$$f_{x^m}(x_{i1}|y_1, X^p, \theta^{*t}) = \frac{g(x_{i1}, y_{i1}, x_i^p, \theta^{*t})}{\int_0^{x_i^p} g(x_{i1}, y_{i1}, x_i^p, \theta^{*t}) dx_{i1}}.$$

We approximate Q using Legendre-Gauss quadrature; the approximation is shown in Supplementary Web Appendix C.

The second step in the EM algorithm is to maximize Q with respect to (wrt) the unknown parameters in θ^* . Notice that β_0^* is a function of κ and the other parameters, $[\beta_1, \beta_2, \beta_3, \gamma']$. Since κ is obtained from previous studies, we maximize Q wrt $[\beta_1, \beta_2, \beta_3, \gamma']$. Define the θ^* which maximizes $Q(\theta^*|\theta^{*t})$ to be $\theta^{*(t+1)}$. We use the optim function in R [15] to maximize Q wrt $[\beta_1, \beta_2, \beta_3, \gamma']$. Once we find $\theta^{*(t+1)}$, we set $t = t + 1$, and repeat the first and second steps of the EM algorithm until there is convergence.

3.2. Variance estimation

To estimate the variance of the parameters estimated by the EM algorithm we used the observed information matrix defined by Louis [16]. Note that β_0 is completely defined by κ and $[\beta_1; \beta_2; \beta_3; \gamma']'$, and κ is defined by the investigator. Define, $\theta^{**} = [\beta_1, \beta_2, \beta_3, \gamma']'$;

therefore, $\theta^* = [\beta_0^*, \theta^{**}]'$, and we wish to estimate the variance of θ^{**} . Louis [16] proposed the following relationship for the observed information matrix for an estimate of θ^{**} , say $\hat{\theta}^{**}$,

$$I_{obs}(\hat{\theta}^{**}) = I_{complete}(\hat{\theta}^{**}) - I_{miss|obs}(\hat{\theta}^{**}),$$

where $I_{complete}(\hat{\theta}^{**})$ is the observed information matrix of the conditional expected complete data, and $I_{miss|obs}(\hat{\theta}^{**})$ is the expected information for the conditional distribution of the exposure variable given the observed data, y_1 and X^p , and an estimate of θ^* , θ^{*t} . The observed information matrix, $I_{complete}(\hat{\theta}^{**})$ is,

$$I_{complete}(\hat{\theta}^{**}) = \frac{-\delta^2 Q(\theta^*|\theta^{*t})}{\delta \theta^{**} \delta \theta^{**t}} \Big|_{\theta^* = \theta^{*t} = [\beta_0^*, \hat{\theta}^{**t}]'}$$

An approximation to the Hessian matrix of Q wrt θ is found in Supplementary Web Appendix D. Additionally, Louis [16] showed that the expected information for the conditional distribution of the exposure variable, $I_{miss|obs}(\hat{\theta}^{**})$, can be simplified as follows,

$$\begin{aligned}
I_{miss|obs}(\hat{\theta}^{**}) &= E_{f_{x^m}} \left\{ \frac{\delta \log [f_{x^m}(X_1, \dots, X_{p-1} | \mathbf{y}_1, X^p, \theta^{*t})]}{\delta \theta^{*t}} \times \frac{\delta \log [f_{x^m}(X_1, \dots, X_{p-1} | \mathbf{y}_1, X^p, \theta^{*t})]'}{\delta \theta^{*t}} \right\} \bigg|_{\theta^{*t} = [\beta_0^*, \hat{\theta}^{*t}]'} \\
&= \sum_{i=1}^{n/p} E_{f_{x^m}} \left\{ \frac{\delta \log [f_{x^m}(\mathbf{x}_{i1}, \dots, \mathbf{x}_{i(p-1)} | \mathbf{y}_1, X^p, \theta^{*t})]}{\delta \theta^{*t}} \times \frac{\delta \log [f_{x^m}(\mathbf{x}_{i1}, \dots, \mathbf{x}_{i(p-1)} | \mathbf{y}_1, X^p, \theta^{*t})]'}{\delta \theta^{*t}} \right\} \bigg|_{\theta^{*t} = [\beta_0^*, \hat{\theta}^{*t}]'}.
\end{aligned}$$

The gradient vector of $\log(f_{x^m})$ wrt θ^{*t} is found in Supplementary Web Appendix E. The expectations, $E_{f_{x^m}}$, can be approximated using Equation S1 found in the Supplementary Web Appendix D.1.

4. Simulation

We implemented simulations to examine the performance of the EM algorithm in a case-control study where biospecimens are measured in pools formed by randomly pairing biospecimen obtained from individuals with the same disease status. All simulations used 1000 independently-generated case-control samples, or Monte Carlo replications. We defined $\beta = [-1, 0.2, 0.5, -0.5]'$ to be approximately equal to the estimated β in our example, shown in Section 5. For Table 1 we defined $\gamma = [0.1, 0.1, 1, 1, 0.5, -2, -2,]'$ to be approximately equal to the parameter estimates obtained in our example.

Table 1 provides the bias, root mean square error (RMSE), observed standard deviation (SD), and estimated SD for β under multiple scenarios. The first scenario individually measures the exposures, e.g. chemokine concentrations, of 5000 biospecimens (2500 cases, e.g. women with miscarriages, and 2500 controls). While individually measuring 5000 biospecimens may provide the ideal amount of information, it may not be practical based on financial limitations. Therefore an investigator may randomly sample 2500 biospecimens (1250 cases and 1250 controls) and individually measure their exposures. The second scenario in Table 1 uses this random sampling approach. As expected this random sampling approach is less efficient than measuring all 5000 available biospecimens (observed SD of the interaction parameter is 0.231 for the 5000 individually measured biospecimens, compared to 0.340 for the 2500 individually measured biospecimens). Rather than randomly sampling 2500 biospecimens to measure, an investigator may choose to pool biospecimens in randomly formed pairs from individuals with the same disease status; thus measure $n^p = 2500$ pools of biospecimens. As discussed in the beginning of Section 3, if an investigator chooses to pool in a case-control study, they may obtain an estimate of the disease prevalence, κ , from a cohort study. The third scenario in Table 1 provides the results using this pooling strategy with a cohort sample size of 20,000 from which κ is estimated. As expected, using all 5000 available biospecimens to measure 2500 pools provides more information than randomly sampling 2500, and provides less information than individually measuring all available biospecimens (i.e. the observed SD of the interaction using pooled biospecimen is 0.234, which is between the observed SD for individually measuring all 5000, 0.231, and individually measuring a random sample of 2500, 0.340). Moreover, note that the percent bias for the traditional logistic regression when $n = 5000$ is 2.4%, and when $n = 2500$ the percent bias is 9.2%. Maximum likelihood estimates are known to be biased in

finite samples, though the bias decreases with increasing sample size. The percent bias for the EM algorithm when biospecimens are measured in pools is 10.2%, which is on par with the traditional logistic regression with $n = 2500$.

To explore the robustness of our methodology to misspecification of the distribution of the pooled covariates, we generated data from independent Gamma distributions with shape and scale parameters 10 and 0.1, respectively. Table 2 provides the bias, RMSE, and observed SD, under the same scenarios as Table 1, with the exception of the misspecified covariate distribution. Table 2 shows similar results to Table 1. The EM algorithm provides relatively unbiased estimators for β . The observed SD of the interaction parameter for the pooled scenario is 0.381, which is between the observed SD for individually measuring all available biospecimen, 0.284, and randomly sampling half, 0.397.

Additionally, two simulation results are provided in the Supplementary Web Appendix G. Table S1 provides the bias, RMSE, observed SD, and estimated SD under the same scenarios as Table 1, with the exception that the sample size for the first scenario is 736, the sample size for the second scenario is 368, and the third scenario observes 368 pools of biospecimen. These sample sizes correspond to the motivating example. Table S1 shows similar results to Table 1. The observed SDs of the interaction parameter for the pooled scenario is 0.800, which is between the observed SD for individually measuring all available biospecimen, 0.718, and randomly sampling half, 1.231. Lastly, Table S2 provides the bias, RMSE, observed SD, estimated SD and type 1 error under the same scenarios as Table 1, with the exception that the interaction parameter, β_3 is defined to be zero. Table S2 shows that the type 1 error is less than the level of the test (0.05). This is likely due to the slight overestimation of the standard deviation, which was also observed in Table 1 and 2.

5. Example

We illustrate the proposed method using a dataset with measurements of macrophage inhibitory protein MIP-1 α and MIP-1 β , for each individual's biospecimen, and for pools of two individuals' biospecimen matched on case status. The individuals selected for this case control study participated in the Collaborative Perinatal Project (CPP) cohort, which consists of approximately 54, 000 pregnant women recruited from 12 sites in the United States between 1959 to 1965. The women selected to participate were matched by gestational age at sample collection, and whether their pregnancy led to a live birth or spontaneous abortion. Details regarding sample collection and demographics are provided elsewhere [9, 11]. Since this study is nested in a cohort study and individuals are matched by gestational age at sample collection, nested case-control study methods may be appropriate. Using the individual level data we estimated the random effect of the matched pair and found that the variance of the random effect was estimated to be zero. Therefore, it is reasonable to analyze our data using traditional case-control study methods rather than nested case-control study methods. However, it is important to note that we were able to test the variance of the random effect because we have access to the individual biospecimen measurements; if an investigators only had access to paired pooled measurements, the variance estimation of the random effect would not be possible. The proposed methods in this manuscript are for a transitional case-control study, future research is necessary to extend the methods to a nested

case-control study. Of the individuals' biospecimen collected for this study, the chemokines MIP-1 α and MIP-1 β were measured twice for 738 samples (360 cases, and 378 controls); once individually, and once within a pool. One individual was excluded from our analysis because their measurement of MIP-1 α was determined to be an outlier (likely due to laboratory measurement error). Therefore, since this individual belonged to a pool of two individuals' biospecimen, our sample size was reduced to 736 samples (358 cases, and 378 controls). The pooled measurements for MIP-1 α and MIP-1 β provide an approximate average chemokine concentration for the individuals in the pool. Our notation defines the pooled exposure, $\mathbf{x}^p$, to be the sum of the individual exposures, therefore, we multiplied all pooled measurements by the pool size, two. This study is one of the only case control studies where both pooled and unpooled samples were collected, making it a valuable resource for validating statistical methodology for analyzing pooled samples.

Due to the unique study design used to obtain this data, we are able to compare the pooled to the traditional individual measurements. Weinberg and Umbach [4, 5] and Schisterman and Vexler [2] claimed that pooling may reduce the percentage of observations below the lower limit of detection (LOD). In fact, we found that 6.39% of individual MIP-1 α , and 5.84% of individual MIP-1 β measurements were below the LOD; compared to 4.89% pooled MIP-1 α and 4.89% pooled MIP-1 β measurements. For simplicity, we replaced measurements below the LOD with 10^{-3} . More sophisticated methods exist for estimating measurements below the LOD, however we analyzed the data with different replacements and found minor differences in the results due to the low percentage of observations below the LOD. Figure 1 plots the log of the sum of the individual measurements versus the log of twice the pooled measurements (recall that observed pooled measurement is an average, therefore we multiply the average by two to obtain the sum). We used the log transformation to appropriately scale the axis in the plots. The horizontal line of observations in each plot are the observations for which the pooled measurements were below the LOD. The square markers are the observations for which at least one of the individual measurements was below the LOD. This figure shows that the agreement between the sum of individual measurements and twice the pooled measurements is very good. In fact, the correlation between the sum of individual measurements and twice the pooled measurements (not log transformed) is 0.95 for MIP-1 α and 0.81 for MIP-1 β . The average absolute difference between the sum of individuals' measurements and twice the pools' measurements is 0.047 pg/mL and 0.048 pg/mL for MIP-1 α and MIP-1 β respectively. The median absolute difference is 0.005 pg/mL and 0.013 pg/mL for MIP-1 α and MIP-1 β respectively. The variance of the individual measurements is 0.129 and 0.027, and the variance of the pooled measurements is 0.062 and 0.017 for MIP-1 α and MIP-1 β respectively. It is important to note that neither measurements are the true concentration levels, rather both measurements (pooled and individual) are subject to measurement error, and the percentage of measurements below the LOD is higher in the individual measurements than the pooled measurements.

When analyzing this dataset, we have the opportunity to compare the parameter estimates from traditional logistic regression using individual measurements, and the proposed EM algorithm using the pooled measurements. We define, $\mathbf{x}_{ij} = [x_{ij1}, x_{ij2}]'$, where x_{ij1} is the

MIP-1 α concentration, and x_{ij2} is the MIP-1 β concentration for the j^{th} person in the i^{th} pool. Therefore, β_0 is the intercept parameter, β_1 is the main effect for MIP-1 α , β_2 is the main effect for MIP-1 β , and β_3 is the interaction effect for MIP-1 $\alpha \times$ MIP-1 β . We use traditional logistic regression to analyze the individual measurements. Furthermore, we took a random sample of half of the individual measurements, and analyzed half of the data using traditional logistic regression (we report the results of a single sample of half of the individual measurements, and the mean of 1000 different random samples). The rationale for taking a random sample of half of the individual measurements is to provide a fair comparison to the analysis of the pooled data. Of course, having full information is very likely to be superior to pooled information. However when an investigator only has enough financial resources to measure half of their sample, the choice is between pooling biospecimens or taking a random sample of one-half the specimens from the biospecimen bank. Estimate of the disease incidence, κ (i.e. the probability of a spontaneous miscarriage) may be obtained from the CPP cohort; however the late gestational age that most women were recruited left many early losses unobserved. Only 830 women (1.53%) of the women recruited in the CPP cohort experienced fetal losses (less than 140 days after the first day of their last menstrual period) [10, 17]. Whitcomb et al. [9] estimated the percentage of pregnancies that end in miscarriage is between 15–31%. Therefore, for our analysis of the pooled biospecimens, we chose $\kappa = 0.3$ and $\kappa = 0.2$ and we compare the results using different estimates of the disease prevalence. Table 3 shows that the interaction effect is negative for all unpooled and pooled designs. Further, the inference on the interaction term is not sensitive to κ since all the estimates were negative and not significant.

6. Discussion

We have shown that the EM algorithm can be a useful tool in estimating an interaction parameter in a logistic regression when biospecimen are measured in pools. We focused on the case-control study, though an area of future research is extending this methodology to prospective studies. Furthermore, we have focused on a simple logistic model with only two main effects and an interaction effect, however it is easy to extend these methods to include more covariates, either measured individually or in pools. Indeed, including additional covariates may improve the efficiency of the pooled design, as noted by the optimal pooling strategies developed by Ma et al.[1]. Moreover, although we have focused on a simple main effects model, the procedures for testing for interaction can be extended to quadratic and higher order main effects model. For example, if X_1 and X_2 are continuous exposures measured in pools, using the EM algorithm approach, a main model could be

$Logit[P(Y=1)] = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + \beta_3 X_2 + \beta_4 X_2^2$, and the interaction model could add the terms $\beta_5 X_1 X_2 + \beta_6 X_1^2 X_2 + \beta_7 X_1 X_2^2 + \beta_8 X_1^2 X_2^2$. These main effect and interaction models can be fit using our proposed methodology.

While the set based logistic model proposed by Weinberg and Umbach [4, 5] is simple and easy to use, it cannot estimate interaction effects when both exposures are measured in biospecimen pools. Our simulations show that randomly sampling half of the available biospecimen has less efficiency than pooling pairs of biospecimens stratified by disease status. These results are encouraging because investigators interested in pooling their

biospecimen to obtain efficiency gains demonstrated by Weinberg and Umbach [4, 5] can now, in addition, apply the proposed method to explore hypotheses relating to interactions or higher order effects using their pooled measurements.

The EM algorithm that we have implemented is similar to a traditional missing data problem due to the assumption that the observed pooled measurements are the sum of unobserved latent individual measurements ($\mathbf{x}^P = \mathbf{x}_1 + \mathbf{x}_2$). Figure 1 shows that this assumption is reasonable. However, when pooling is used to reduce cost or the quantity of required biospecimen, the only exposure data that is observed and measured is $\mathbf{x}^P$. Throughout this manuscript we have formulated the likelihood with $\mathbf{x}_2 = \mathbf{x}^P - \mathbf{x}_1$ and $\mathbf{x}_1$ is unknown. It is important to note that the results are identical if we were to formulate the problem as $\mathbf{x}_1 = \mathbf{x}^P - \mathbf{x}_2$ with $\mathbf{x}_2$ unknown. Furthermore, fifty iterations were used for each EM chain in simulation. Visual inspection of the EM chains showed that convergence was relatively fast (by iteration 20–30). In simulations not provided, we ran the simulations with 100 iterations and found the results did not change. In practice, one should monitor parameter estimates from iteration to iteration until convergence.

It is important to note that we do not explicitly account for measurement error, and this may be an explanation for the difference between the results obtained from the traditional logistic regression and the proposed EM algorithm. Table 3 showed that the point estimates of the interaction using the pooled information (−0.563 and −0.470) were in the same direction as the point estimate using the full individual information (−1.405), and both were found to be not statistically significant. In the CPP data example, the pooled measurement is not exactly the sum of the two individual measurements, as noted in Figure 1. However, the individual measurements and the pooled measurement are similar, the correlation between the sum of individual measurements and pooled measurements was very strong, 0.95 for MIP-1 α and 0.81 for MIP-1 β . Measurement error occurs both in measuring the biospecimens individually, and when measuring them in pools, therefore we would not expect that the sum of the two individual measurements to be exactly the same as the pooled measurement. Moreover, we showed that pooling reduced the number of measurements below the LOD, which agreed with claims made by Weinberg and Umbach [4, 5] and Schisterman and Vexler [2], and suggests that measurement error may be improved when measuring pools of biospecimens. Exploring measurement error is a very interesting topic for future research. A method to do this may require some replication at the individual (non-pooled) level.

Additionally, the EM algorithm requires specifying a marginal distribution of exposures which are measured in pools, and an estimate of the disease prevalence, κ . We believe that the Box-Cox transformation is useful because of its flexibility and ability to approximate exposure distributions which are usually skewed (note, the Box-Cox transformation was purely used to find a flexible distribution of the exposures, it did not effect the interpretation of the parameters in the logistic model). When appropriate, the EM algorithm can be modified to handle other parametric marginal distributions for the exposures, f_x . In simulations not shown, we considered independent Gamma distributions for the marginal distribution of the exposures, f_x . While estimation improves as the number of unknown parameters decreases (independent Gamma distributions for the marginal distribution of the exposures has four unknown parameters compared to the Box-Cox transformation

distribution which has seven unknown parameters), if the assumption of independence between the exposures is wrong we found that assuming independent Gamma distributions will yield very biased estimates of the logistic model parameters. Further research is necessary to explore options to relax or test the assumption for the marginal distribution of exposures, and the estimate of the disease prevalence, κ .

Our simulations showed that estimating the variance of the EM estimator using the observed information matrix [16] was not robust to incorrect choices of κ nor misspecified marginal distribution of the exposures. Further research is necessary to explore options for estimating the variance of the EM estimators. The Monte Carlo EM (MCEM) may improve the performance of the observed covariance matrix if the approximation of the integrals performed poorly. Additionally, other methods such as bootstrap [18] and profile likelihood confidence intervals may have better performance [19].

The EM algorithm was implemented for pool sizes of two, since our motivating chemokine example was designed in pool sizes of two. The pool size may be larger than two, though when the pool size increases the dimensionality of the required integration also increases, and numerical approximations of the high dimensional integral may perform poorly. The R code we provide in the online supplement is specific for $p = 2$. Other methods, such as MCEM may be appropriate for larger pool sizes. Efficiency and robustness of pooling for interaction testing will likely depend heavily on pool size. Intuitively the estimation of interaction will become more difficult as the pool size increases. Furthermore, the pool size need not be constant for the entire sample. Some samples may be measured individually, while others are measured in pools of size two, and others are measured in pools of size three, and so forth. Mixing pool sizes may yield an optimal pooling strategy with regards to resources and estimation efficiency [3]. Further research needs to explore the optimal pooling strategies while using the EM algorithm for parameter estimation.

Additionally, we assumed that pools are randomly formed within disease status, which was the situation in our motivating example. Therefore once the disease status is known, pools are formed by randomly pairing biospecimen that were obtained from individuals with the same disease status. Pooling within disease status and within strata, for example pooling within age group, is an interesting extension to this research. Intuitively, pooling specimens that are expected to have similar values will improve the performance of the EM algorithm by decreasing the variance within a pool. However, the likelihood would need to account for pooling within strata of a covariate, resulting in an increased amount of complexity for implementation. Further research needs to explore the benefits of pooling within disease status and covariate strata.

Lastly, this manuscript described methods for data obtained in a case-control study. As discussed in Section 5 the data in our motivating example was obtained in a nested case-control study with individuals matched on gestational age at sample collection. However, the selected cytokine levels were not related to the time of entry into the study, and using the individual level data we estimated the variance was zero for the random effect of the matched pair, therefore using traditional case-control study methods was appropriate.

However, extending the proposed methods to handling matching in a nested case-control study is an interesting proposal for future research.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This research was supported by the Long-Range Research Initiative of the American Chemistry Council and the Intramural Research Program of the *Eunice Kennedy Shriver* National Institute of Child Health and Human Development (NICHD), National Institutes of Health, Bethesda, MD (Contracts # HHSN275200403394C, HHSN2752011000021 Task 1 HHSN27500001). This work utilized the computational resources of the NIH HPC Biowulf cluster. (<http://hpc.nih.gov>)

References

1. Ma CX, Vexler A, Schisterman EF, Tian L. Cost-efficient designs based on linearly associated biomarkers. *Journal of Applied Statistics*. 2011; 38:2739–2750.10.1080/02664763.2011.567254
2. Schisterman EF, Vexler A. To pool or not to pool, from whether to when: applications of pooling to biospecimens subject to a limit of detection. *Paediatric and Perinatal Epidemiology*. 2008; 22:486–496.10.1111/j.1365-3016.2008.00956.x [PubMed: 18782255]
3. Schisterman EF, Vexler A, Mumford SL, Perkins NJ. Hybrid pooled-unpooled design for cost-efficient measurement of biomarkers. *Statistics in Medicine*. 2010; 29:597–613.10.1002/sim.382 [PubMed: 20049693]
4. Weinberg CR, Umbach DM. Using pooled exposure assessment to improve efficiency in case-control studies. *Biometrics*. 1999; 55:718–726.10.1111/j.0006-341X.1999.00718.x [PubMed: 11314998]
5. Weinberg CR, Umbach DM. Correction to “Using pooled exposure assessment to improve efficiency in case-control studies”, by Clarice R. Weinberg and David M. Umbach;55, 718–726, September 1999. *Biometrics*. 2014; 70:1061.10.1111/biom.12261
6. Zhang Z, Albert PS. Binary regression analysis with pooled exposure measurements: A regression calibration approach. *Biometrics*. 2011; 67:636–645.10.1111/j.1541-0420.2010.01464.x [PubMed: 20662830]
7. Mumford SL, Schisterman EF, Vexler A, Liu A. Pooling biospecimens and limits of detection: effects on ROC curve analysis. *Biostatistics*. 2006; 7:585–598.10.1093/biostatistics/kxj027 [PubMed: 16531470]
8. Laird SM, Tuckerman E, Cork BA, Linjawi S, Blakemore AIF, Li TC. A review of immune cells and molecules in women with recurrent miscarriage. *Human Reproduction Update*. 2003; 9:163–174.10.1093/humupd/dmg013 [PubMed: 12751778]
9. Whitcomb BW, Schisterman EF, Klebanoff MA, Baumgarten M, Rhoton-Vlasak A, Luo X, Chagini N. Circulating chemokine levels and miscarriage. *American Journal of Epidemiology*. 2007; 166:323–331.10.1093/aje/kwm084 [PubMed: 17504778]
10. Keim S, Klebanoff M. Aspirin use and miscarriage risk. *Epidemiology*. 2006; 17:435–439. [PubMed: 16755260]
11. Whitcomb BW, Perkins NJ, Zhang Z, Ye A, Lyles RH. Assessment of skewed exposure in case-control studies with pooling. *Statistics in Medicine*. 2012; 31:2461–2472.10.1002/sim.5351 [PubMed: 22437722]
12. Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society Series B (Methodological)*. 1977; 39:1–38.
13. Box G, Cox D. An analysis of transformations. *Journal of the Royal Statistical Society Series B (Methodological)*. 1964; 26:211–252.
14. Prentice RL, Pyke R. Logistic disease incidence models and case-control studies. *Biometrika*. 1979; 66:403–411.10.1093/biomet/66.3.403

15. R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing; Vienna, Austria: 2014. <http://www.R-project.org/>
16. Louis TA. Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society Series B (Methodological)*. 1982; 44:226–233.
17. Klebanoff M, Levine R, DerSimonian R, Clemens J, Wilkins DG. Maternal serum paraxanthine, a caffeine metabolite, and the risk of spontaneous abortion. *The New England Journal of Medicine*. 1999; 341:1639– 1644. [PubMed: 10572151]
18. Efron B, Gong G. A leisurely look at the bootstrap, the jackknife, and cross-validation. *The American Statistician*. 1983; 37:36–48.10.1080/00031305.1983.10483087
19. Venzon DJ, Moolgavkar SH. A method for computing profile-likelihood-based confidence intervals. *Journal of the Royal Statistical Society Series C (Applied Statistics)*. 1988; 37:87–94.

Figure 1.

Log transformation of the sum of individual measurements versus the pooled measurement times two. The square marker indicate the pools where at least one of the individual measurements was below the LOD.

Table 1

Simulation results based on 1000 Monte Carlo replications. The data were generated with the true parameters, $\beta = [-1, 0.2, 0.5, -0.5]'$, and $\gamma = [0.1, 0.1, 1, 1, 0.5, -2, -2]'$. A case control design was employed for all simulations with $n/2$ samples in each of the case and control groups. An estimate of the disease prevalence, κ , was obtained from a cohort with 20,000 samples. Estimated SD¹ is the standard deviation estimated using the method proposed by Louis (1982) as discussed in Section 3.2

	$\beta_1 = 0.2$	$\beta_2 = 0.5$	$\beta_3 = 0.5$
Traditional logistic regression, $n = 5000$			
Bias	0.001	-0.011	0.012
RMSE	0.143	0.145	0.232
Observed SD	0.143	0.145	0.231
Estimated SD	0.141	0.142	0.224
95% Confidence Interval Coverage	0.950	0.953	0.945
Traditional logistic regression, $n = 2500$			
Bias	-0.005	-0.021	0.046
RMSE	0.209	0.215	0.343
Observed SD	0.209	0.214	0.340
Estimated SD	0.203	0.205	0.332
95% Confidence Interval Coverage	0.953	0.942	0.956
EM algorithm, $n = 5000, p = 2$			
Bias	0.035	0.010	-0.051
RMSE	0.158	0.149	0.239
Observed SD	0.154	0.149	0.234
Estimated SD ¹	0.151	0.151	0.253
95% Confidence Interval Coverage	0.932	0.948	0.950

Table 2

Simulation results based on 1000 Monte Carlo replications. The data were generated with the true parameters, $\beta = [-1, 0.2, 0.5, -0.5]'$, and the marginal distribution of the covariates was defined by independent Gamma distributions with shape and scale parameters 10 and 0.1, respectively. A case control design was employed for all simulations with $n/2$ samples in each of the case and control groups. An estimate of the disease prevalence, κ , was obtained from a cohort with 20,000 samples. Estimated SD¹ is the standard deviation estimated using the method proposed by Louis (1982) as discussed in Section 3.2

	$\beta_1 = 0.2$	$\beta_2 = 0.5$	$\beta_3 = 0.5$
Traditional logistic regression, $n = 5000$			
Bias	0.016	0.009	-0.015
RMSE	0.299	0.295	0.285
Observed SD	0.299	0.295	0.284
Estimated SD	0.301	0.300	0.288
95% Confidence Interval Coverage	0.950	0.959	0.951
Traditional logistic regression, $n = 2500$			
Bias	-0.024	-0.022	0.020
RMSE	0.413	0.414	0.398
Observed SD	0.412	0.414	0.397
Estimated SD	0.427	0.426	0.410
95% Confidence Interval Coverage	0.957	0.958	0.953
EM algorithm, $n = 5000, p = 2$			
Bias	0.061	0.062	-0.061
RMSE	0.395	0.389	0.386
Observed SD	0.390	0.384	0.381
Estimated SD ¹	0.419	0.417	0.411
95% Confidence Interval Coverage	0.961	0.958	0.960

Table 3

Parameter estimates in the CPP application comparing women with and without miscarriages. Here β_1 is the main effect for MIP-1 α , β_2 is the main effect for MIP-1 β , and β_3 is the interaction effect for MIP-1 $\alpha \times$ MIP-1 β . The quantity SD^1 is the standard deviation estimated using the method proposed by Louis (1982) as discussed in Section 3.2.

	β_1	β_2	β_3
Traditional logistic regression, $n = 736$			
Estimate	0.217	1.943	-1.405
SD	0.628	1.012	0.849
p-value	0.730	0.055	0.098
Traditional logistic regression, $n = 368$			
Estimate (one random sample)	0.049	3.971	-2.888
SD (one random sample)	0.940	1.706	1.937
p-value (one random sample)	0.959	0.020	0.134
Estimate (mean of 1000 random samples)	0.490	2.198	-2.106
SD (mean of 1000 random samples)	1.063	1.545	1.666
p-value (mean of 1000 random samples)	0.557	0.233	0.288
EM algorithm, $n = 736, p = 2, \kappa = 0.3$			
Estimate	-0.121	-0.283	-0.563
SD^1	0.647	1.121	1.396
p-value	0.885	0.801	0.687
EM algorithm, $n = 736, p = 2, \kappa = 0.2$			
Estimate	-0.393	-0.207	-0.470
SD^1	0.571	0.850	0.709
p-value	0.491	0.808	0.507