Published in final edited form as:

Biometrics. 2014 March ; 70(1): 202–211. doi:10.1111/biom.12134.

Regression for Skewed Biomarker Outcomes Subject to Pooling

Emily M. Mitchell^{1,*}, Robert H. Lyles¹, Amita K. Manatunga¹, Michelle Danaher², Neil J. Perkins², and Enrique F. Schisterman²

¹Department of Biostatistics and Bioinformatics, Rollins School of Public Health, Emory University, Atlanta, Georgia 30322, U.S.A.

²Division of Intramural Population Health Research, Eunice Kennedy Shriver National Institute of Child Health and Human Development, Bethesda, Maryland 20892, U.S.A.

Summary

Epidemiological studies involving biomarkers are often hindered by prohibitively expensive laboratory tests. Strategically pooling specimens prior to performing these lab assays has been shown to effectively reduce cost with minimal information loss in a logistic regression setting. When the goal is to perform regression with a continuous biomarker as the outcome, regression analysis of pooled specimens may not be straightforward, particularly if the outcome is right-skewed. In such cases, we demonstrate that a slight modification of a standard multiple linear regression model for poolwise data can provide valid and precise coefficient estimates when pools are formed by combining biospecimens from subjects with identical covariate values. When these x-homogeneous pools cannot be formed, we propose a Monte Carlo Expectation Maximization (MCEM) algorithm to compute maximum likelihood estimates (MLEs). Simulation studies demonstrate that these analytical methods provide essentially unbiased estimates of coefficient parameters as well as their standard errors when appropriate assumptions are met. Furthermore, we show how one can utilize the fully observed covariate data to inform the pooling strategy, yielding a high level of statistical efficiency at a fraction of the total lab cost.

Keywords

Biomarkers; Design; Efficiency; MCEM; Pooled specimens; Skewness

1. Introduction

Physically combining specimens into composite samples, or pools, prior to performing laboratory assays has various advantages. Traditionally, pooling has been employed to reduce cost by decreasing the total number of diagnostic tests required to detect disease (Dorfman, 1943) and is often standard procedure when testing for HIV in donated blood (Lan et al., 1993; HHS, 2010). In epidemiological studies, Weinberg and Umbach (1999)

*emitch8@emory.edu.

Supplementary Materials

Supplementary Web Appendices and example code referenced in Sections 4, 5.2, 5.3, 6.2, 7.1, and 9 are available with this paper at the *Biometrics* website on Wiley Online Library.

and Vansteelandt et al. (2000) have demonstrated the advantages of pooling in a logistic regression setting in order to reduce cost while preserving irreplaceable specimens, and Schisterman and Vexler (2008) illustrate the potential reduction in information loss by pooling in the presence of a detection limit. Yet only recently has the extension of pooling methods to applications involving regression with continuous outcome variables been considered (Ma et al., 2011; Malinovsky et al., 2012), and its potential cost-saving benefits in this arena have remained largely untapped.

When faced with budgetary constraints for laboratory assays, a subset of biospecimens may be randomly selected for analysis commensurate with available funds. This practice, however, can result in a detrimental loss of information compared with the potential of the entire collection of specimens. Alternatively, combining available specimens into pools based on subject-specific characteristics offers the potential to both maintain a high level of statistical precision and remain within financial confines. Such strategic pooling techniques have been shown to provide efficiency gains over random selection in logistic regression settings when pooling is applied to a binary outcome or exposure of interest (Weinberg and Umbach, 1999; Vansteelandt et al., 2000; Zhang and Albert, 2011; Zhang et al., 2012; Lyles et al., 2012), or to an untransformed outcome in linear regression (Ma et al., 2011). Many biomarkers assessed by laboratory assay techniques, however, lead to positive and right-skewed continuous measurements that require a transformation when treated as the dependent variable in regression analyses. In such cases, care must be taken when performing regression on transformed poolwise as opposed to individual measurements, as doing so can produce invalid coefficient estimates.

In this study, we explore situations for which standard regression techniques applied to pooled outcome data are defensible, and outline conditions under which more complicated analytical methods are required. In the latter case, we propose a specific Monte Carlo Expectation Maximization (MCEM) algorithm to calculate maximum likelihood estimates (MLEs) of the regression coefficients. In the following sections, we introduce the motivating dataset for this study and then describe the analytical methods under consideration, along with the conditions required for their validity. Section 6 summarizes various pooling methods, with particular focus on a k -means clustering-based approach that promotes estimate efficiency. In Section 7, we provide simulation studies that test the proposed analytical strategies and illustrate the benefits of informative pooling. Finally, we apply these methods to our motivating example.

2. A Motivating Example: Cytokines in the CPP

The Collaborative Perinatal Project (CPP) was conducted between 1959 and 1974 to examine associations between various exposures and pregnancy outcomes (Hardy, 2003). In a nested case-control study of stored serum samples from the CPP, several cytokines were measured in participants that experienced a spontaneous abortion (SA), along with controls matched to cases by gestational age (GA) at sample collection (Whitcomb et al., 2007). Accompanying covariates included participant demographics such as age, race, and smoking status. While the cytokines in relation to SA from this CPP study have previously been analyzed via logistic regression with SA status as the dependent variable (Whitcomb et al.,

2008, 2012), our study treats monocyte chemotactic protein 1 (MCP1) as the outcome, and SA status, age, race, and smoking status as predictors. The positive, right-skewed nature of MCP1, as well as nearly all the cytokines measured in this study, motivates the development of methods to analyze pooled, skewed outcomes in a regression setting. Specifically, we seek to estimate the parameters of an underlying individual-level lognormal regression model for the dependent variable MCP1, when measurements of MCP1 are obtained on pooled samples.

This dataset is particularly compelling since it contains both individual-level as well as pooled measurements of the cytokines, where pools were formed randomly within SA status (maximum pool size = 2) as part of a study design incorporating a methods component. This unique characteristic enables analysis of the effect of pooling on parameter estimation compared with estimates calculated from measurements on the individual specimens. While pools in this dataset were formed homogeneously on SA status, later in this paper we discuss additional design strategies that utilize all covariate information to improve efficiency. We use this dataset collected from the CPP study to illustrate analytical methods and to demonstrate the advantages of informative pooling techniques, so that future studies of this type can benefit from the increased precision provided by these strategic designs.

3. Regression Model for Individual Subjects

Performing linear regression on a right-skewed biomarker often invites a log transformation. Suppose that the log of the outcome is linearly associated with the predictor variables, so that the true model can be represented by:

$$\log(Y_{ij}) = \alpha + \mathbf{x}_{ij}\beta + \epsilon_{ij}, \quad \mathbf{j} = 1, \dots, \mathbf{k}_i, \quad \mathbf{i} = 1, \dots, \mathbf{n}, \quad (1)$$

where α is the intercept, β is the $P \times 1$ column vector of coefficients, and Y_{ij} , ϵ_{ij} , and $\mathbf{x}_{ij} = (x_{ij1}, \dots, x_{ijP})$ are the outcome, error, and row vector of covariates for the j^{th} subject in the i^{th} pool, respectively. Furthermore, let $N = \sum_{i=1}^n k_i$ denote the total number of subjects, where k_i represents the number of specimens in pool i (i.e., pool size). The ϵ_{ij} 's are assumed independent and identically distributed with $E(\epsilon_{ij}) = 0$ and $Var(\epsilon_{ij}) = \sigma^2$. If the value of each individual's outcome were known, a straightforward application of multiple linear regression (MLR) on the log-transformed outcome would yield the desired parameter estimates. Similarly, if n individual specimens are selected for analysis, the same MLR estimation procedure could be applied to this subset of the full data. When specimens are pooled, however, only the measured value of the pool is known, while each specimen's outcome (Y_{ij}) remains unobserved, so that a simple application of (1) to the pooled measurements might not be appropriate. Details on pooling strategies are described in Section 6. For now, we consider methods for analyzing data based on specimens that have already been pooled, where our objective is valid and efficient estimation of the vector β that appears in (1).

4. Least Squares Regression on Pooled Outcomes

A natural inclination when faced with analyzing pooled, right-skewed data may be to perform linear regression on a log-transformation of the measured values of each pool:

$$\text{Naive Model: } \log(Y_i^p) = \alpha + \bar{\mathbf{x}}_{i\bullet} \beta + \delta_i^{(1)},$$

where $\bar{\mathbf{x}}_{i\bullet} = (\bar{x}_{i\bullet 1}, \dots, \bar{x}_{i\bullet p})$ is the pooled vector of predictors such that $\bar{x}_{i\bullet p} = \frac{1}{k_i} \sum_{j=1}^{k_i} x_{ijp}$ is the arithmetic mean of the p^{th} predictor across all specimens in pool i , and the measured

outcome of pool i , $Y_i^p = \frac{1}{k_i} \sum_{j=1}^{k_i} Y_{ij}$, is assumed to reflect the arithmetic mean of the individual concentrations constituting that pool. This naive model will often produce biased estimates of the regression coefficients, due to the non-linearity of the log function. Yet while $\log(Y_i^p)$ may not be defined by the model assumptions, its expectation (conditional on $\mathbf{X}$) can be approximated by a second-order Taylor series expansion, so that

$$E\{\log(Y_i^p)\} \approx \log\left\{\frac{1}{k_i} \sum_{j=1}^{k_i} E(Y_{ij})\right\} - \frac{\sum_{j=1}^{k_i} \text{Var}(Y_{ij})}{2\left\{\sum_{j=1}^{k_i} E(Y_{ij})\right\}^2}.$$

When pools are x -homogeneous, such that $\mathbf{x}_{ij} = \bar{\mathbf{x}}_{i\bullet}$ for all $j = 1, \dots, k_i$, this reduces to:

$$E\{\log(Y_i^p)\} \approx \alpha + \bar{\mathbf{x}}_{i\bullet} \beta + \log(a) - \frac{c}{2k_i}, \quad (2)$$

where $a = E(e^{\epsilon^{11}})$ and $c = \text{Var}(e^{\epsilon^{11}})/E(e^{\epsilon^{11}})^2$. Since the expectation of these log-transformed pools are a function of pool size (k_i) we propose the following approximate model to reduce any bias associated with applying the transformation:

$$\text{Approximate Model: } \log(Y_i^p) = \alpha + \gamma k_i^{-1} + \bar{\mathbf{x}}_{i\bullet} \beta + \delta_i^{(2)},$$

where γ is the regression coefficient corresponding to k_i^{-1} and $\delta_i^{(2)}$ represents the error term for pool i under this model, where we are still working under the assumption of x -homogeneous pools, i.e. $\mathbf{x}_{ij} = \bar{\mathbf{x}}_{i\bullet}$ for all $j = 1, \dots, k_i$. A similar approximation can be applied to estimate $\text{Var}\{\log(Y_i^p)\}$, where, under a first-order Taylor series expansion, $\text{Var}\{\log(Y_i^p)\} \approx c/k_i$, suggesting that a weighted least squares (WLS) analysis with weight matrix $\mathbf{K} = \text{diag}(k_1, \dots, k_n)$ may be beneficial in improving estimate precision. The preceding $\text{diag}()$ notation represents the diagonal matrix with (i, i) element equal to k_i ($i = 1, \dots, n$). Note that when all pool sizes are equal, this model reduces to the Naive Model.

Let $\mathbf{X}_2 = (\mathbf{1} \ \mathbf{K}^{-1} \ \mathbf{X})$. Then the WLS estimate for $\boldsymbol{\theta} = (\alpha, \gamma, \beta')'$ under the Approximate Model is $\hat{\boldsymbol{\theta}} = (\mathbf{X}_2' \mathbf{K} \mathbf{X}_2)^{-1} \mathbf{X}_2' \mathbf{K} \log(\mathbf{Y}^p)$. Now, applying the approximation in (2), we get:

$$E(\hat{\theta}) = E \begin{pmatrix} \hat{\alpha} \\ \hat{\gamma} \\ \hat{\beta} \end{pmatrix} \approx (\mathbf{X}_2' \mathbf{K} \mathbf{X}_2)^{-1} \mathbf{X}_2' \mathbf{K} \mathbf{X}_2 \begin{pmatrix} \alpha + \log(a) \\ -c/2 \\ \beta \end{pmatrix} = \begin{pmatrix} \alpha + \log(a) \\ -c/2 \\ \beta \end{pmatrix}.$$

Under this model, $\hat{\alpha}$ is biased by a factor of $\log(a)$, $\hat{\gamma}$ is an approximately unbiased estimator of $-c/2$, and $\hat{\beta}$ will be an approximately unbiased estimator of the original coefficient vector β . The estimated variance of $\hat{\theta}$, $\widehat{Var}(\hat{\theta}) = \hat{c}(\mathbf{X}_2' \mathbf{K} \mathbf{X}_2)^{-1}$, is approximately unbiased as well, where $\hat{c}$ is the usual WLS variance estimate (see Supplementary Web Appendix C for details). When the total number of pools (n) is large, $\hat{\theta}$ will be approximately normally distributed due to asymptotic properties under the central limit theorem, so that the usual 95% confidence intervals based on the normal distribution should provide nominal 95% coverage in large samples. Since this property only applies when n is large, applying the standard t reference distribution with $n - P - 1$ degrees of freedom is a reasonable measure to help alleviate overly liberal confidence intervals when sample size is small.

One advantage of analyzing homogeneous pools under the Approximate Model is that fully-specified distributional assumptions are not required, since the validity of this method relies only on the correct specification of the first two moments characterizing the individual-level specimens. In Section 7, we demonstrate the potential repercussions of assuming the Naive Model and the advantages of applying the Approximate Model to analyze x-homogeneous pools. The simplicity of the Approximate Model as well as its flexibility in not requiring any specific distributional assumptions are bolstered by simulation results.

5. Calculating MLEs

It is not always possible to form x-homogeneous pools, especially if one or more of the covariates are continuous. In such cases, the Taylor series approximations from Section 4 are no longer justified. Instead, parametric approaches to identify MLEs of the β vector may be the best option. While these methods do require distributional assumptions, they provide theoretically sound alternatives to the Approximate Model when pools are heterogeneous.

A natural method to calculate MLEs is to maximize the observed data likelihood directly. For pooled specimens, the density for pool i is characterized by the $(k_i - 1)$ -fold integral

$$f_p(Y_i^p | \mathbf{X}, \theta) = \int_{Y_{ik_i}} \cdots \int_{Y_{i2}} k_i I \left(\sum_{j=2}^{k_i} Y_{ij} < k_i Y_i^p \right) \left\{ f_1 \left(k_i Y_i^p - \sum_{j=2}^{k_i} Y_{ij} \right) \prod_{j=2}^{k_i} f_j(Y_{ij}) \right\} dY_{i2} \cdots dY_{ik_i}, \quad (3)$$

where $I(\bullet)$ is the indicator function, and $f_j(y) = f(y | \mathbf{x}_{ij}, \theta)$ is the assumed density of the individual level data for the j^{th} subject in pool i that depends on the parameter vector θ as well as the covariate vector $\mathbf{x}_{ij}$. When there are at most two specimens in each pool (i.e. $k_i \leq 2$ for all i), the observed likelihood can often be maximized through existing numerical integration and optimization functions such as the *integrate* and *optim* functions in R or the QUAD and NLPQN procedures in SAS IML. For larger pool sizes, however, numerical optimization of the likelihood can quickly become computationally intractable. The

integrand characterizing the density of a sum of lognormal random variables, in particular, has a reputation for being especially poorly-behaved (Beaulieu and Xie, 2004; Santos Filho et al., 2006). In subsequent simulations and analyses, we apply direct optimization via the Convolution formula (3) when possible. For larger pools sizes (i.e. $k_i \geq 3$) we propose a Monte Carlo Expectation Maximization (MCEM) algorithm as a more dependable tool to optimize the observed likelihood.

5.1 MCEM Algorithm

In lieu of maximizing a complicated observed log-likelihood, the EM algorithm seeks to maximize the expected value of the conditional log-likelihood, which is often more accessible when the complete data are assumed to follow a distribution from which MLE calculation is straightforward (Dempster et al., 1977; Wei and Tanner, 1990). The first step in the EM algorithm (the “E Step”) calculates the expectation of the complete log-likelihood

given the observed data. Let $g(Y_{i2}, \dots, Y_{ik_i} | Y_i^p, \mathbf{X}, \theta^{(t)})$ denote the density of the missing data (i.e. individual-level measurements) given the observed data (i.e. pooled measurements) under the parameter vector $\theta^{(t)}$ and fully observed covariate data $\mathbf{X}$. Without loss of generality, Y_{i1} is excluded from the missing data since its value is fixed given Y_i^p and

$(Y_{i2}, \dots, Y_{ik_i})$, due to the restriction $\sum_{j=2}^{k_i} Y_{ij} = k_i Y_i^p$. The expected conditional log-likelihood is then:

$$Q(\theta | \theta^{(t)}) = E \left\{ \log L_c(\theta) | Y^p, \mathbf{X}, \theta^{(t)} \right\} = \sum_{i=1}^n E_g \left[\log f \left\{ \left(k_i Y_i^p - \sum_{j=2}^{k_i} Y_{ij} \right) | \mathbf{x}_{i1}, \theta \right\} + \sum_{j=2}^{k_i} \log f(Y_{ij} | \mathbf{x}_{ij}, \theta) \right] \quad (4)$$

where $\theta^{(t)} = (\alpha^{(t)}, \beta'^{(t)}, \sigma'^{(t)})'$ is the estimate of the parameter vector at the t^{th} iteration of the algorithm. Let $h(Y_{i[-1]})$ represent any of the continuous functions of the missing data contained in (4), where $\mathbf{Y}_{i[-1]} = (Y_{i2}, \dots, Y_{ik_i})$. For a right-skewed distribution (e.g. lognormal), finding a closed form expression for $E_g\{h(\mathbf{Y}_{i[-1]})\}$ can be difficult. In such cases, we apply Monte Carlo methods to approximate this value.

5.2 Monte Carlo Estimation

By the weak law of large numbers (WLLN), $E_g\{h(\mathbf{Y}_{i[-1]})\}$ can be estimated by:

$$E_g \left\{ h(\mathbf{Y}_{i[-1]}) \right\} = \int_{Y_{ik_i}} \dots \int_{Y_{i2}} h(\mathbf{Y}_{i[-1]}) g(\mathbf{Y}_{i[-1]} | Y_i^p, \mathbf{X}, \theta^{(t)}) d\mathbf{Y}_{i[-1]} \approx \frac{1}{M} \sum_{m=1}^M h(\mathbf{Y}_{i[-1],m}),$$

where $\mathbf{Y}_{i[-1],m} = (Y_{i2,m}, \dots, Y_{ik_i,m})$ is generated under the joint conditional distribution $g(\mathbf{Y}_{i[-1]} | Y_i^p, \mathbf{X}, \theta^{(t)})$ for each m , and M is a number large enough for the asymptotic properties of the WLLN to hold. Several strategies for choosing the best values of M at each iteration have been explored (Booth and Hobert, 1999; Levine and Casella, 2001; Wei and Tanner, 1990). For simulations presented in this paper, we speed convergence by calculating starting values under the Approximate Model, which will often target the appropriate neighborhood of the coefficient estimates. the iteration process begins with $M = 50$, and

after every 20 iterations, M is increased by 25%. The algorithm then runs for 500 iterations, since additional iterations produce only negligible changes in the parameter estimates.

Applying Bayes' Rule, we can re-write $g(\mathbf{Y}_{i[-1]}|Y_i^p, \mathbf{X}, \theta^{(t)})$ as :

$$g(\mathbf{Y}_{i[-1]}|Y_i^p, \mathbf{X}, \theta^{(t)}) = \frac{k_i I\left(\sum_{j=2}^{k_i} Y_{ij} < k_i Y_i^p\right) f\left(k_i Y_i^p - \sum_{j=2}^{k_i} Y_{ij} | \mathbf{x}_{i1}, \theta^{(t)}\right) \prod_{j=2}^{k_i} f(Y_{ij} | \mathbf{x}_{ij}, \theta^{(t)})}{f_p(Y_i^p | \mathbf{X}, \theta^{(t)})}$$

where we have incorporated the linear restriction on the Y_{ij} 's into the density expression as an indicator function. The main difficulty in generating data from g is meeting this inequality constraint contained in the indicator function. While a rejection sampling technique could be employed to generate $Y_{i[-1],m}$ under this restriction, a more computationally efficient method can be achieved through importance sampling (Lange, 2010). The basic idea behind this strategy is to identify a distribution that is similar to g , but from which samples are easier to obtain. Importance weights are then applied to the Monte Carlo estimate of $E_g\{h(\mathbf{Y}_{i[-1]})\}$ in order to adjust for generating data under the alternate distribution:

$$E_g\{h(\mathbf{Y}_{i[-1]})\} \approx \frac{\sum_{m=1}^M h(\mathbf{Y}_{i[-1],m}) w(\mathbf{Y}_{i[-1],m})}{\sum_{m=1}^M w(\mathbf{Y}_{i[-1],m})} \quad (5)$$

where $w(\mathbf{Y}_{i[-1]}) = g/g^*$ and each $\mathbf{Y}_{i[-1],m} = (Y_{i2,m}, \dots, Y_{ik_i,m})$ is now generated under the alternate distribution, $g^*(\mathbf{Y}_{i[-1]}|Y_i^p, \mathbf{X}, \theta^{(t)})$ (details in Supplementary Web Appendix C. 3).

Since the linear constraint $\sum_{j=2}^{k_i} Y_{ij,m} < k_i Y_i^p$ poses the main difficulty, we can choose the alternate distribution to satisfy this restriction first. While there may be multiple candidates

for $g^*(\mathbf{Y}_{i[-1]}|Y_i^p, \mathbf{X}, \theta^{(t)})$, one straightforward choice is to first generate each $Y_{ij,m}$ from $f(Y_{ij}|\mathbf{x}_{ij}, \theta^{(t)})$ for $j = 1, \dots, k_i$, then alter each $Y_{ij,m}$ to form the sample $Z_{i,m} = (Z_{i1,m}, \dots,$

$Z_{ik_i,m})$, where $Z_{ij,m} = Y_{ij,m} (k_i Y_i^p) / \left(\sum_{j=1}^{k_i} Y_{ij,m}\right)$, which meets the linear constraint. To calculate importance weights, we must first determine the appropriate expression for

$g^*(\mathbf{Z}_{i[-1],m}|Y_i^p, \mathbf{X}, \theta^{(t)})$. Following the derivation outlined in Frigyi et al. (2010), the joint density of $(Z_{i1,m}, \dots, Z_{ik_i,m})$ can be found by applying the change of variable:

$$(Y_{i1,m}, \dots, Y_{ik_i,m}) = \left\{ S \left(k_i Y_i^p - \sum_{j=2}^{k_i} Z_{ij,m} \right), SZ_{i2,m}, \dots, SZ_{ik_i,m} \right\},$$

where S is defined as $S = \left(\sum_{j=1}^{k_i} Y_{ij,m}\right) / k_i Y_i^p$. The joint distribution of $(S, \mathbf{Z}_{i[-1],m})$ is then:

$$g_s \left(S, \mathbf{Z}_{i[-1],m} | Y_i^p, \mathbf{X}, \theta^{(t)} \right) = |J| \times f \left\{ S \left(k_i Y_i^p - \sum_{j=2}^{k_i} Z_{ij,m} \right) | Y_i^p, \mathbf{x}_{i1}, \theta^{(t)} \right\} \prod_{j=2}^{k_i} f \left(S Z_{ij,m} | Y_i^p, \mathbf{x}_{ij}, \theta^{(t)} \right), \quad (6)$$

where $|J| = k_i Y_i^p S^{k_i-1}$ is the determinant of the Jacobian. Integrating over the domain of S will give the expression for $g^* \left(\mathbf{Z}_{i[-1],m} | Y_i^p, \mathbf{X}, \theta^{(t)} \right)$. When this expression has a closed form, calculation of the importance weights is straightforward.

Once the conditional expectations have been approximated, maximizing Q with respect to θ (the “M Step”) is a fairly simple task. This is particularly true when the assumed distribution of the outcome is a member of the exponential family, for which update formulas are often simple to maximize numerically using existing software (e.g. SAS, R).

5.3 Standard Error Estimation

One of the drawbacks of an EM type algorithm is that calculating standard error estimates can prove difficult. For this study, since Monte Carlo techniques are required to approximate the conditional expectations of the MLEs, we apply MC approximations to strategies similar to Louis’s method to estimate the observed information matrix (Louis, 1982). The Hessian of the observed likelihood can be written as:

$$\sum_{i=1}^n \left[\frac{d^2 Q_i}{d\theta^2} - \left(\frac{dQ_i}{d\theta} \right) \left(\frac{dQ_i}{d\theta} \right)' + E_g \left\{ \left(\frac{d}{d\theta} \log f_c \left(Y_i^p, \mathbf{Y}_{i[-1]} | \mathbf{X}, \theta \right) \right) \left(\frac{d}{d\theta} \log f_c \left(Y_i^p, \mathbf{Y}_{i[-1]} | \mathbf{X}, \theta \right) \right)' \right\} \right], \quad (7)$$

where $f_c \left(Y_i^p, \mathbf{Y}_{i[-1]} | \mathbf{X}, \theta \right)$ is the density of the complete data for pool i (derivation in Supplementary Web Appendix C.4). Monte Carlo methods can be used to approximate each component of the summation in (7), which can then be inverted to give the negative of the variance-covariance matrix of the MLEs.

While these MCEM methods could theoretically be applied to any parametric model, we provide explicit steps for a lognormal model for pooled data in Supplementary Web Appendix B, as this distribution is commonly applied to right-skewed outcomes in regression settings.

6. Pooling Methods

For this study, we assume that N specimens have been collected, but only n ($< N$) lab tests can be ordered. As mentioned previously, randomly selecting n of the available specimens allows for straightforward regression analysis but often results in a considerable loss of efficiency. An alternative approach to reduce the number of lab tests while maintaining a high level of efficiency incorporates covariate data into the pooling process. So long as pools are based only on the fully observed covariate values, appropriately defined regression coefficient estimates and estimated standard errors remain valid (Little and Rubin, 2002).

6.1 *x-homogeneous Pools*

As demonstrated in Section 4, when pools are formed from subjects with identical covariate values, valid coefficient estimates can easily be obtained by applying the Approximate Model to log-transformed pooled values. In order to form *x-homogeneous* pools, the number of desired pools (n) must be larger than the number of unique groups with identical covariate values, say G . For this study, homogeneous pools were formed under the following conditions:

1. Each group of unique covariates is required to supply at least one pool.
2. Groups containing only one member contribute that individual specimen for analysis.
3. When possible, pools are formed to have similar sizes.

6.2 *k-means Clustering*

When it is not possible to form *x-homogeneous* pools (e.g. when $n < G$ or at least one covariate is continuous), a *k-means* clustering algorithm can target a pooling strategy that minimizes the within-pool sum of squares (Hartigan, 1975). The resulting clusters consist of observations with similar covariates values, helping to preserve the relationship between the covariates and outcome while often maintaining much of the precision from the full dataset.

This algorithm is readily accessible in software packages such as the *kmeans* function in R and the FASTCLUS procedure in SAS (Example R code is provided in the Supplementary Materials). The function accepts values for the data (here, the matrix of fully-known covariates) as well as the desired number of clusters, making it a natural clustering mechanism when the number of available lab assays is fixed by budgetary constraints. When all predictors are considered equally important, standardizing each variable prior to performing *k-means* clustering will ensure that each variable contributes similar influence on the resulting clusters. The *k-means* algorithm will often (but not always) form clusters that are homogeneous with respect to any binary or categorical variables. In recent work (Mitchell et al., unpublished manuscript), the *k-means* algorithm has proved particularly effective at maintaining high precision levels when multiple linear regression is applied to a pooled outcome.

7. Simulation Study

For each of the following scenarios, 5000 simulations were performed in R. Datasets were generated to resemble actual motivating data described in Section 2, with sample size $N = 672$. Independent predictor variables were generated to mimic age (years), smoking status (yes/no), race (1 = white / 2 = black), and SA status (yes/no), and the outcome variable to resemble the cytokine MCP1 ($\mu\text{g/mL}$) based on a lognormal regression against those predictors. Age was simulated as a normal random variable with mean 26.6 and standard deviation 6.4, then rounded to the nearest whole number (this permits the formation of *x-homogeneous* pools when average pool size is small). Smoking status, race, and SA status were simulated as Bernoulli random variables with probabilities 0.47, 0.28, and 0.46, respectively. The outcome, MCP1, was generated under a lognormal distribution such that

$E\{\log(\text{MCP1})|X\} = -2.48 + 0.017(\text{Age}) + 0.007(\text{Smoking Status}) - 0.388(\text{Race}) + 0.132(\text{SA})$ and $\text{Var}\{\log(\text{MCP1})|X\} = 1.19$. In the first study, we assess the performance of each of the proposed analytical strategies on x-homogeneous pools ($n = 336$), and in the next study we compare a k -means pooling strategy to random pooling and selection, when x-homogeneous pools cannot be formed ($n = 112$).

7.1 Comparing Analytical Strategies

The goal of this first simulation study is to assess each of the discussed analytical strategies for x-homogeneous pools, where analysis is deemed appropriate if it provides accurate estimates of the regression coefficients as well as their standard errors. Pools were formed based on the x-homogeneous clustering strategy described in Section 6.1, where pool sizes ranged from 1 to 6. Analytical strategies under consideration include standard least squares regression on log-transformed pooled outcomes (Naive Model), WLS on the log-transformed pools with inverted pool size as a predictor variable (Approximate Model), and the likelihood-based MCEM strategy under lognormal regression (MCEM Model). We also provide regression results from the full data as well as a random sample of size $n = 336$ for comparison purposes. Since many of the pools in this simulation contained more than 2 specimens, direct optimization of the likelihood under the Convolution approach was not viable. Additional simulations comparing the Convolution approach to the MCEM and Approximate models are available in Supplementary Web Appendix A.

Table 1 displays the mean bias and empirical standard deviation (SD) of the regression coefficient estimates. The ratio of mean estimated standard error to empirical standard deviation ($\widehat{SE}/SD$) is also provided, where a value of 1 is ideal. 95% confidence interval (CI) coverage is based on the estimated standard errors and a t -reference distribution with $n - 5$ degrees of freedom.

Estimates from the Naive Model tend to be biased, which can result in severe CI under-coverage, most notably with a coverage rate of less than 81% for $\hat{\beta}_3$. The remaining methods provide approximately unbiased estimates of the regression coefficients (Mean Bias ≈ 0) as well as their estimated standard errors ($\widehat{SE}/SD \approx 1$) and close to 95% CI coverage. These results support the validity of the proposed Approximate and MCEM Models as appropriate analytical methods for x-homogeneous pools.

Although the main purpose of this simulation is to test the performance of the proposed methods, it is also worth noting that estimates from these x-homogeneous pools analyzed under the Approximate and MCEM Models are noticeably more precise than those from a random sample, and are only slightly less efficient than those from the full dataset. The MCEM method appears to provide marginally more precise estimates than those under the Approximate Model, most likely due to the fact that this method identifies MLEs, which are well-known to be the most efficient estimates when the assumed underlying distribution is correctly specified, as in this simulation. This slight improvement in efficiency, however, is unlikely to motivate the additional computational time and effort required to implement the MCEM method. In addition, while the MCEM method requires a fully specified distribution on the individual-level errors, the Approximate Model only requires assumptions on the first

two moments (1). Thus, the Approximate Model may be the most desirable analytical method when pools are x-homogeneous, due to its straightforward application and its flexibility as a semi-parametric model.

7.2 Comparing Pooling Strategies

In the next simulation, we compare regression results based on k -means vs. random pooling, when x-homogeneous pools cannot be formed due to a small total number of pools ($n = 112$). k -means clustering was performed using the *kmeans* function in R, where pool sizes ranged from 1 to 49, with an average size of 6. Each of the pooled strategies was analyzed under the MCEM algorithm, since the heterogeneity of the pools and large pool sizes precluded defensible application of the Approximate Model and Convolution method.

Based on the results in Table 2, all pooling and selection strategies provide approximately unbiased estimates with close to nominal 95% CI coverage. While random pooling seems to give slightly biased estimates of β_3 , further examination suggests that this apparent bias is a consequence of the estimate exhibiting a slightly left-skewed distribution, likely due to the sample size being too small for asymptotic normality of this MLE to apply. Estimates of β_3 under k -means pooling, on the other hand, do not exhibit this characteristic, suggesting that asymptotic properties may apply for a smaller number of pools when k -means clustering is performed. k -means pooling also provides coefficient estimates that are considerably more precise than both random strategies, more than doubling the efficiency for each of the estimates, and losing surprisingly little precision relative to the full data analysis even at $1/6^{\text{th}}$ the original sample size. Thus, a considerable amount of information can be retained from the full data when k -means pooling is performed and appropriate analytical techniques are applied, at just a fraction of the total laboratory cost.

8. Data Analysis

Analysis of the CPP substudy example was conducted on data from 672 participants who provided complete information on MCP1 as well as each of the 4 covariates. The single MCP1 measurement that fell below the detection limit was assigned a value of 0.00001. In addition, specimens from 508 of the participants had been combined to form 254 actual pools, each containing 2 specimens (Whitcomb et al., 2012). MCP1 values were measured again on these pools. Thus, we have access to MCP1 measurements from the complete dataset (672 lab assays), as well as from a dataset of 254 pools and 164 individual specimens (418 lab assays). For this analysis, we first perform linear regression on a log-transformation of MCP1 on the 672 individual measurements. These results serve as the “gold standard” for subsequent analyses, since they represent the maximum information available from the dataset. Next, we analyze the set of 418 observed pooled measurements and individual specimens. As an additional comparison, we then perform regression on the same set of 418 pools and individuals, but this time, we use artificial pooling to determine the expected value of each pool, calculated as the arithmetic mean of the measurements from individual specimens. These observed and expected pooled datasets are analyzed under both the MCEM algorithm and the Convolution method.

For our final analysis, x-homogeneous pools are created artificially, in order to illustrate the results that would have been available had the entire set of observed covariate information been used to inform the pooling process. At the desired sample size ($n = 418$) pool sizes ranged from 1 to 5 in order to maintain homogeneity among the pooled covariates. The Convolution approach was not available due to these larger pool sizes. Instead, we analyze these x-homogeneous pools using the MCEM algorithm and the Approximate Model.

Results of these data analyses are provided in Table 3. Estimates from the MCEM and Convolution approaches are almost identical, validating the performance of the MCEM model as an appropriate algorithm for estimating MLEs. The observed and expected pools provide similar conclusions with respect to estimated standard errors and significance levels, although the actual estimates tend to vary. This discrepancy is likely due to measurement error or pooling error, a topic explored in depth by Schisterman et al. (2010) when measuring an exposure of interest on pooled samples. While detailed evaluation of these potential sources of error is beyond the scope of the current study, this issue highlights the importance of addressing the potential effects of additional error components when analyzing biomarkers.

For x-homogeneous pools, the MCEM algorithm and the Approximate Model provide almost identical results, emphasizing the advantages of the more accessible Approximate Model. Estimates from these artificially-formed pools are generally most similar to those from the full data analysis, and all are more precise than estimates obtained from the actual pools, which were formed homogeneously only with respect to SA status (Whitcomb et al., 2012). In addition, regression estimates from these x-homogeneous pools concur with the full data results that race is significantly associated with levels of MCP1, a relationship that is lost when pools are randomly formed with respect to this covariate. As demonstrated by this data analysis and corroborated by the simulation studies, utilizing the entire covariate information to create pools can preserve associations present in the full dataset, maintain high levels of efficiency, and even simplify the analytical process when pools are x-homogeneous.

8.1 Regression Diagnostics

One concern when considering a pooling strategy is the potential difficulty in assessing model assumptions. When the Approximate Model is applied to x-homogeneous pools, the entire set of pooled values can contribute to diagnostic procedures, since regression assumptions are made on approximations of the first two moments of the pooled data. Thus, graphical illustrations including all pooled observations can indicate whether these assumptions are violated. For other models, however, we advocate a hybrid pooling approach which includes measurements on both pooled as well as individual specimens. That way, assuming enough individual specimens are available, these can be checked for any indication of a violation of the model assumptions. Often, the k -means algorithm will automatically single out individual specimens. It may be of interest, however, to verify that enough individual measurements will be available to properly identify any anomalies prior to performing any physical pooling.

Figure 1 provides an illustration of residual plots available under each of the pooling strategies. Each graph compares standardized Pearson residuals to fitted values, where the residual for the i^{th} specimen or pool (when applicable) is defined as

$$r_i = (\log y_i - \hat{\mu}_i) \left\{ \widehat{Var}(\log Y_i) (1 - \hat{h}_i) \right\}^{-1/2},$$

with $\hat{\mu}_i$ and $\widehat{Var}(\log Y_i)$ denoting the model-specific estimates of the mean and variance of $\log y_i$, respectively, and $\hat{h}_i$ the leverage (Agresti, 2002). The hybrid pooling design from the data analysis (corresponding to Observed and Expected pools) permits graphical analysis of a subset of the individual-level measurements only, since it is not straightforward to define the residuals for non-homogeneous pools. Although the artificially-formed x-homogeneous pools also benefit from a hybrid design, the application of the Approximate Model enables use of all available measurements, since the residuals are based on a weighted least squares analysis applied to the pooled values. In all cases, the observation corresponding to ID 823 is noted as a potential outlier. In fact, this was the subject whose measured MCP1 value fell below the detection limit, to which we applied the value of 0.00001. While these diagnostic plots are available for pooled data so long as a sufficient number of individual specimens are available, this illustration emphasizes the advantages of the Approximate Model in its ability to use all pooled values in model assessment. Further diagnostic considerations are the subject of future work (see Discussion).

9. Discussion

Our goal for this paper was to develop methods for analyzing pooled, right-skewed data, specifically when a log-transformation is needed on the individual-level outcome and budgetary constraints limit the number of assays that can be performed. When covariate data is available prior to any physical pooling, this information can be utilized to form pools that will produce precise regression estimates, often losing a minimal amount of information even at a fraction of the original sample size. When possible, forming x-homogeneous pools will not only tend to maximize efficiency for a particular sample size, but can also permit exploitation of a Taylor series approximation, so that a suitably specified linear regression model applied to the log of the observed, pooled values will yield appropriate and precise estimates. If specimens have already been pooled, as in the motivating dataset from the CPP study, or if it is not possible to form x-homogeneous pools, MLEs of regression estimates can be calculated via a Convolution method when pool sizes do not exceed 2, or by an MCEM algorithm that can accommodate larger pool sizes.

While we have focused our discussion on regression under a lognormal distribution, additional distributions may be similarly suitable for modeling positive, right-skewed data. In particular, the gamma distribution has unique properties that retain a closed form of the likelihood under x-homogeneous pools, which has previously been applied to pooled data by Whitcomb et al. (2012). While it is possible to use only the unpooled data to assess model assumptions, as demonstrated in Figure 1, much of the original information may be lost. These graphical diagnostics may help identify violations of the model assumptions, but are unlikely to help determine whether a lognormal or gamma distribution, for instance, would provide a better fit. Ongoing research is exploring extensions of analytical indicators of model fit to pooled data, in order to better assess model assumptions by incorporating all

available information. In Supplementary Web Appendix A, we provide additional simulation studies to illustrate the potential consequences of distributional and link misspecification on the proposed analytical methods.

As demonstrated by the data analysis, actual datasets are rarely immune to issues such as measurement or pooling error, limit of detection, or missing covariate values. While these topics have previously been considered when estimating the mean and variance of a pooled exposure of interest (Schisterman and Vexler, 2008; Schisterman et al., 2010, 2011), current research is working toward extending these methods to the regression settings considered here. In particular, Schisterman et al. (2011) show that a hybrid pooled-unpooled design can be effective in evaluating pooling and measurement error for a gamma or normal distribution. We anticipate that a similar design could prove helpful in evaluating error components in a regression setting, further emphasizing the benefit of a hybrid pooling design. When covariate data is missing, exclusion of individuals with missing information will not affect the validity of the resulting estimates, so long as missingness depends only on the covariates (Little and Rubin, 2002). Doing so, however, is unlikely to be the most efficient mitigation strategy. Another way to deal with this issue is to force any observations containing missing data to be analyzed as individual specimens, which may facilitate application of standard missing data techniques. The pooling techniques discussed in this paper are currently unable to handle observations with missing data. Additional investigation of improved strategies for incorporating these types of data points into a pooling process invites further research.

Naturally, we want to reduce cost as much as possible, in this scenario, by reducing the number of lab assays. Of course, even when strategic pooling methods are employed, any reduction in lab assays will often correspond to a reduction in precision. Additional simulations (see Supplementary Web Appendix A) indicate that when the total number of pools is small, the proposed analytical methods still provide virtually unbiased estimates of the regression coefficients, but standard error estimates tend to suffer, suggesting that the total number of pools should be large enough for asymptotic properties to hold. Ultimately, the best pooling design will depend on many additional considerations specific to each study. A thorough evaluation of cost savings vs. potential precision reduction, feasibility and practicality of implementation, and inclusion of individual specimens to facilitate assessment of model fit and potential measurement or pooling error is highly recommended. Pooling can be a valuable tool in reducing the cost of lab assays, and the methods discussed in this paper provide a foundation for performing regression analyses on a right-skewed, pooled outcome. Regardless, the advantages and disadvantages of any pooling strategies, whenever possible, should be carefully considered prior to implementation.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This research was supported by grants from the National Institute of Nursing Research (1RC4NR012527-01) and from the National Institute of Environmental Health Sciences (5R01ES012458-07). Additional partial support from

the National Center for Advancing Translational Sciences of the National Institutes of Health under Award Number UL1TR-000454 is also acknowledged. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

- Agresti, A. *Categorical Data Analysis*. Wiley; Hoboken: 2002.
- Beaulieu NC, Xie Q. An Optimal Lognormal Approximation to Lognormal Sum Distributions. *IEEE Transactions on Vehicular Technology*. 2004; 53:479–489.
- Booth JG, Hobert JP. Maximizing Generalized Linear Mixed Model Likelihoods with an Automated Monte Carlo EM Algorithm. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*. 1999; 61:265–285.
- 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.
- Dorfman R. The detection of defective members of large populations. *Annals of Mathematical Statistics*. 1943; 14:436–440.
- Frigyik, BA.; Kapila, A.; Gupta, MR. UWEE Technical Report UWEETR-2010-0006. Department of Electrical Engineering, University of Washington; 2010. Introduction to the Dirichlet Distribution and Related Processes..
- Hardy JB. The Collaborative Perinatal Project: lessons and legacy. *Annals of Epidemiology*. 2003; 5:303–311. [PubMed: 12821268]
- Hartigan, JA. *Clustering Algorithms*. Wiley; New York: 1975.
- HHS, FDA, Center for Biologics Evaluation and Research. Guidance for Industry, Nucleic Acid Testing for Human Immunodeficiency Virus Type 1 and Hepatitis C Virus (HCV): Testing, Product Disposition, and Donor Deferral and Reentry. 2010
- Lan S-J, Hsieh C-C, Yen Y-Y. Pooling Strategies for Screening Blood in Areas with Low Prevalence of HIV. *Biometrical Journal*. 1993; 35:553–565.
- Lange, K. *Numerical Analysis for Statisticians*. 2nd edition. Springer; New York: 2010.
- Levine RA, Casella G. Implementations of the Monte Carlo EM Algorithm. *Journal of Computational and Graphical Statistics*. 2001; 10:422–439.
- Little, RJA.; Rubin, DB. *Statistical Analysis with Missing Data*. Second Edition. Wiley; New York: 2002.
- 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.
- Lyles RH, Tang L, Lin J, Zhang Z, Mukherjee B. Likelihood-based methods for regression analysis with binary exposure status assessed by pooling. *Statistics in Medicine*. 2012; 31:2485–2497. [PubMed: 22415630]
- Ma C-X, Vexler A, Schisterman EF, Tian L. Cost-efficient designs based on linearly associated biomarkers. *Journal of Applied Statistics*. 2011; 38:2739–2750.
- Malinovsky Y, Albert PS, Schisterman EF. Pooling Designs for Outcomes under a Gaussian Random Effects Model. *Biometrics*. 2012; 68:45–52. [PubMed: 21981372]
- Santos Filho JCS, Yacoub MD, Cardieri P. Highly accurate range-adaptive lognormal approximation to lognormal sum distributions. *Electronics Letters*. 2006; 42
- 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. *Pediatric and Perinatal Epidemiology*. 2008; 22:486–496. [PubMed: 18782255]
- 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. [PubMed: 20049693]
- Schisterman EF, Vexler A, Yi A, Perkins NJ. A combined efficient design for biomarker data subject to a limit of detection due to measuring instrument sensitivity. *The Annals of Applied Statistics*. 2011; 5:2651–2667.
- Vansteelandt S, Goetghebeur E, Verstraeten T. Regression models for disease prevalence with diagnostic tests on pools of serum samples. *Biometrics*. 2000; 56:1126–1133. [PubMed: 11129470]

- Wei GCG, Tanner MA. A Monte Carlo Implementation of the EM Algorithm and the Poor Man's Data Augmentation Algorithms. *Journal of the American Statistical Association*. 1990; 85:699–704.
- Weinberg C, Umbach D. Using pooled exposure assessment to improve efficiency in case-control studies. *Biometrics*. 1999; 55:718–726. [PubMed: 11314998]
- 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. [PubMed: 17504778]
- Whitcomb BW, Schisterman EF, Klebanoff MA, Baumgarten M, Luo X, Chagini N. Circulating levels of cytokines during pregnancy: thrombopoietin is elevated in miscarriage. *Fertility and Sterility*. 2008; 89:1795–1802. [PubMed: 17706203]
- 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. [PubMed: 22437722]
- Zhang Z, Albert PS. Binary regression analysis with pooled exposure measurements: a regression calibration approach. *Biometrics*. 2011; 67:636–645. [PubMed: 20662830]
- Zhang Z, Liu A, Lyles RH, Mukherjee B. Logistic regression analysis of biomarker data subject to pooling and dichotomization. *Statistics in Medicine*. 2012; 31:2473–2484. [PubMed: 21953741]

Figure 1.

Plot of fitted values vs. standardized Pearson residuals under full data, x-homogeneous pools analyzed under the Approximate Model, and those corresponding to individual specimens from models based on Observed and Expected pools. Although measured values of these individual specimens are the same for Observed and Expected pools, residuals and fitted values differ slightly, due to their dependency on the model-specific parameter estimates.

Table 1

Simulation results for regression analysis from various analytical models on x-homogeneous pools. “SD” refers to empirical standard deviation of regression coefficient estimates and “ $\widehat{SE}$ ” is the mean estimated standard error.

Mean Bias (SD)					
$\widehat{SE}/SD$ (95% CI Coverage)					
Method	n	$\beta_1 = 0.017$	$\beta_2 = 0.007$	$\beta_3 = -0.388$	$\beta_4 = 0.132$
Full Data	672	0.000 (0.007)	-0.002 (0.085)	-0.001 (0.092)	0.001 (0.085)
		0.99 (94.8)	1.00 (95.1)	1.02 (95.4)	1.00 (94.7)
Random Sample	336	0.000 (0.009)	-0.003 (0.120)	-0.003 (0.130)	0.000 (0.118)
		1.00 (95.1)	1.00 (94.6)	1.03 (95.7)	1.02 (95.3)
Naive Model	336	0.000 (0.007)	-0.016 (0.099)	-0.111 (0.106)	-0.016 (0.099)
		0.89 (92.0)	1.00 (94.8)	0.98 (80.9)	1.01 (94.6)
Approximate Model	336	0.000 (0.007)	-0.002 (0.092)	-0.002 (0.105)	0.001 (0.092)
		0.99 (94.6)	0.98 (94.7)	1.00 (95.3)	0.98 (94.3)
MCEM Model	336	0.000 (0.007)	-0.002 (0.091)	-0.003 (0.098)	0.001 (0.090)
		0.98 (94.3)	0.99 (94.9)	1.01 (95.0)	1.00 (94.8)

Table 2

Simulation results for regression analysis on various pooling methods using the MCEM algorithm. “SD” refers to empirical standard deviation of regression coefficient estimates and “ $\widehat{SE}$ ” is the mean estimated standard error.

Mean Bias (SD)					
$\widehat{SE}/SD$ (95% CI Coverage)					
Method	n	$\beta_1 = 0.017$	$\beta_2 = 0.007$	$\beta_3 = -0.388$	$\beta_4 = 0.132$
Full Data	672	0.000 (0.007)	-0.002 (0.085)	-0.001 (0.092)	0.001 (0.085)
		0.99 (94.8)	1.00 (95.1)	1.02 (95.4)	1.00 (94.7)
Random Sample	112	0.000 (0.017)	-0.002 (0.210)	0.002 (0.231)	0.004 (0.212)
		0.99 (95.2)	1.00 (95.1)	1.01 (95.6)	0.99 (95.1)
Random Pools	112	0.000 (0.019)	-0.009 (0.247)	-0.018 (0.318)	-0.001 (0.251)
		0.96 (93.8)	0.97 (94.6)	0.97 (95.4)	0.95 (94.2)
<i>k</i> -means Pools	112	0.000 (0.008)	-0.003 (0.101)	-0.005 (0.108)	0.000 (0.099)
		0.96 (93.8)	0.97 (94.5)	0.99 (95.1)	0.98 (94.5)

Table 3

Results from regression analyses on the individual and pooled dataset.

Estimate (SE)					
Data	Age (years)	Smoking Status (yes/no)	Race (black/white)	SA Status (yes/no)	$\hat{\sigma}$
Full Data	0.017 (0.007)*	0.007 (0.085)	−0.388 (0.095)*	0.132 (0.086)	1.09
Observed Pools					
MCEM	0.020 (0.009)*	0.083 (0.127)	−0.159 (0.146)	0.135 (0.104)	1.23
Convolution	0.020 (0.010)*	0.083 (0.127)	−0.159 (0.148)	0.136 (0.104)	1.23
Expected Pools					
MCEM	0.021 (0.009)*	0.045 (0.125)	−0.214 (0.146)	0.122 (0.101)	1.21
Convolution	0.021 (0.009)*	0.045 (0.124)	−0.214 (0.146)	0.122 (0.101)	1.21
Homogeneous Pools					
MCEM	0.017 (0.007)*	0.026 (0.092)	−0.306 (0.102)*	0.143 (0.093)	1.12
Approx.	0.016 (0.007)*	0.022 (0.092)	−0.308 (0.103)*	0.141 (0.092)	1.17

A “*” indicates predictors that were found to be significantly associated with log(MCP1) at the 0.05 level. “Full Data” was analyzed via linear regression on log(MCP1), “Observed Pools” and “Expected Pools” under MCEM and the Convolution method, and “Homogeneous Pools” under both MCEM and the Approximate Model.