Published in final edited form as:
Stat Med. 2017 July 10; 36(15): 2363–2377. doi:10.1002/sim.7291.

A general framework for the regression analysis of pooled biomarker assessments

Yan Liu[a], **Christopher S. McMahan**[a,*,†], and **Colin M. Gallagher**[a]

[a]Department of Mathematical Sciences, Clemson University, Clemson, SC 29634, U.S.A

Abstract

As a cost efficient data collection mechanism, the process of assaying pooled biospecimens is becoming increasingly common in epidemiological research; e.g. pooling has been proposed for the purpose of evaluating the diagnostic efficacy of biological markers (biomarkers). To this end, several authors have proposed techniques that allow for the analysis of continuous pooled biomarker assessments. Regretfully, most of these techniques proceed under restrictive assumptions, are unable to account for the effects of measurement error, and fail to control for confounding variables. These limitations are understandably attributable to the complex structure that is inherent to measurements taken on pooled specimens. Consequently, in order to provide practitioners with the tools necessary to accurately and efficiently analyze pooled biomarker assessments, herein a general Monte Carlo maximum likelihood based procedure is presented. The proposed approach allows for the regression analysis of pooled data under practically all parametric models and can be used to directly account for the effects of measurement error. Through simulation, it is shown that the proposed approach can accurately and efficiently estimate all unknown parameters and is more computational efficient than existing techniques. This new methodology is further illustrated using monocyte chemotactic protein-1 data collected by the Collaborative Perinatal Project in an effort to assess the relationship between this chemokine and the risk of miscarriage.

Keywords

Biomarker pooling; Measurement error; Monte Carlo maximum likelihood estimation; Pooled biospecimens

1. Introduction

In resource limited environments, the process of assaying pooled biospecimens (i.e., a sample comprised of several individual specimens) has become a cost effective alternative to assaying specimens one-by-one. The origins of group (or pool) testing are commonly attributed to [1], where it was proposed as a means to reduce the cost of screening military inductees for syphilis during the Second World War. Since its advent, pool testing has been

[*]Correspondence to: Christopher S. McMahan, O-110 Martin Hall, Department of Mathematical Sciences, Clemson University, Clemson, SC 29634, U.S.A.
[†]mcmaha2@clemson.edu

adopted for the purposes of screening for various infectious diseases [2, 3, 4, 5] and the incidence of bioterrorism [6], identifying lead compounds in drug discovery [7], and detecting rare mutations in genetics [8]. The general advantages of collecting data on pools are three-fold: a reduction in the cost associated with measuring the outcome of interest, the ability to preserve irreplaceable specimens, and the means to collect information in a more timely fashion. Further, these advantages persist when the outcome of interest is continuous. For example, [9] reports that the 2005–2006 National Health and Nutrition survey reduced the number of analytical measurements required to characterize the distribution of polychlorinated and polybrominated compounds within the population of the United States from 2201 to 228 by pooling, which translated to a savings of $2.78 million in testing cost.

[10] was the first to consider modeling outcomes obtained from assaying pooled specimens in order to estimate population level characteristics; i.e., binary outcomes, measured on pools, were used to estimate the proportion of individuals within a population who possessed a characteristic of interest. In this context, the process of pooling offers a cost effective data collection mechanism, and for this reason has received a great deal of attention among the statistical literature; e.g., see [11, 12] and the references therein. Extending this earlier work, [13] proposed a regression framework which relates binary outcomes measured on pools to covariate information. The work of [13] has since seen numerous generalizations; to include allowing for random pooling and imperfect testing [14], confirmatory testing [15], random effects [16], and covariate measurement error [17], as well as the development of techniques that allow for a nonparametric [18, 19] and semiparametric [20, 21] regression analysis of group testing data. Note, all of the aforementioned techniques were specifically developed for binary outcomes measured on pools.

Broadening the utilitarian nature of pooling as a cost effective data collection mechanism, techniques for analyzing continuous outcomes obtained from assaying pooled specimens have been proposed. For example, in an effort to reduce cost, several authors have proposed methods of analyzing pooled measurements in order to evaluate the efficacy of a biological marker (biomarker) as a diagnostic tool; e.g., see [22, 23, 24, 25, 26, 27, 28]. More recently, this research area has shifted to consider the regression analysis of continuous biomarker measurements taken on pools; e.g., see [29] and Malinovsky et al. (2012). It is worthwhile to point out that the aforementioned regression techniques were developed under the rather stringent assumption that the biomarker levels of the individuals, and hence the pools, are conditionally Gaussian, given the covariate information. In most practical applications, biomarker levels tend to follow a right-skewed continuous distributions with positive support. Consequently, [30] proposed an Monte Carlo expectation maximization (MCEM) algorithm which could be used to conduct the regression analysis of pooled biomarker assessments under the assumption that the individual biomarker levels conditionally, given the covariates, follow a log-normal distribution. Further, these authors investigated several pooling strategies with respect to estimation efficiency, with their findings resounding the work of [31] in the binary regression group testing literature; i.e., measurements taken on pools that are formed homogeneously, with respect to covariate information, can be used to construct estimators that are nearly as efficient as the analogous estimators based on individual level data. Regretfully, the technique proposed by these authors allows for the regression analysis under a single parametric model and does not account for measurement error, which is omnipresent in biomarker evaluation studies.

The regression analysis of continuous outcomes measured on pools is fraught with many complexities, thus the potential benefits from using pooling as a cost efficient data collection mechanism has been largely untapped when the response variable of interest is continuous. To circumvent this hurdle, herein a general regression methodology for continuous pooled biomarker assessments is proposed. Unlike previously proposed techniques, this methodology allows for the regression analysis under many common parametric models, to include distributions belonging to the class of generalized linear models, and can easily account for measurement error in the response variable, when it is present. Further, the proposed technique is more computationally efficient than other existing methods; e.g., MCEM algorithm of [30]. The asymptotic properties of the proposed approach are established, and through simulation studies the new methodology is shown to accurately and efficiently analyze pool response data, both subject to and free of measurement error, under several different parametric models.

The remainder of this article is organized as follows. Section 2 presents the general modeling framework which can be used to perform a regression analysis of continuous outcomes measured on pooled specimens. The asymptotic properties of the proposed approach are provided in Section 3, and Section 4 provides a simulation study which investigates the finite sample performance of the new methodology. In Section 5 the proposed approach is used to analyze monocyte chemotactic protein-1 data collected by the Collaborative Prenatal Project (CPP). Section 6 concludes with a summary discussion. All of the theoretical proofs and additional technical details are provided in the Web Appendix.

2. Methodology

In what follows a general methodology is proposed for the regression analysis of continuous outcomes measured on pooled specimens. In this context the observed data consists of J measurements taken on pools, where the jth pool is formed by amalgamating c_j specimens collected from individuals. Let $\tilde{Y}_{ij}$ denote the biomarker level of the ith individual in the jth pool, for $i = 1, \ldots, c_j$ and $j = 1, \ldots, J$. When assessments are being made on pools the $\tilde{Y}_{ij}$ are latent, and the observed data consists of either the biomarker level of the pool, which is denoted by $\tilde{Y}_{pj}$, or an error contaminated measurement of $\tilde{Y}_{pj}$, which is denoted by Y_{pj}. In order to relate the $\tilde{Y}_{ij}$ to the $\tilde{Y}_{pj}$, herein it is assumed that the biomarker level of the jth pool is the arithmetic average of the biomarker levels of the individuals of which the jth pool is comprised; i.e., $\tilde{Y}_{p_j}=c_j^{-1}\sum_{i=1}^{c_j}\tilde{Y}_{ij}$ This assumption is common among the literature [26, 27, 30] and is reasonable as long as pools are formed from specimens of equal volume.

For modeling purposes, it is assumed that $\mathbf{x}_{ij} = (1, x_{ij1}, \ldots, x_{ijp})'$, a $(p + 1) \times 1$ vector of covariates, is available for each individual. Given the covariate information, it is assumed that the biomarker levels of the individuals are (conditionally) independent and follow a continuous distribution with probability density function $f(\cdot|\mathbf{x}_{ij}, \boldsymbol{\theta}_0)$; i.e., $\tilde{Y}_{ij} \overset{ind.}{\sim} f(\cdot|\mathbf{x}_{ij},\boldsymbol{\theta}_0)$, for $i = 1, \ldots, c_j$ and $j = 1, \ldots, J$, where $\boldsymbol{\theta}_0=(\boldsymbol{\beta}_0',\boldsymbol{\gamma}_0')'$ denotes the collection of model parameters, $\boldsymbol{\beta}_0 = (\beta_{00}, \ldots, \beta_{0p})'$ is a vector of regression coefficients, and $\boldsymbol{\gamma}_0$ is a set of nuisance parameters. The generality in the assumed parametric model for $\tilde{Y}_{ij}$ is meant to illustrate the broad applicability of the proposed approach; e.g., $f(\cdot|\mathbf{x}_{ij}, \boldsymbol{\theta}_0)$ could belong to the class of generalized linear models, with $\tilde{Y}_{ij}$ being related to $\mathbf{x}_{ij}$ in the usual fashion [32], or the regression analysis could be performed under other common parametric models for biomarker data, such as the Weibull and log-normal distributions; e.g., see [30].

If the individual biomarker levels were observed, the maximum likelihood estimator (MLE) of $\boldsymbol{\theta}_0$ could be obtained through traditional techniques once the parametric model is assumed. However, when pooled assessments are being made, the $\tilde{Y}_{ij}$ are latent and the observed data consists of either $\tilde{Y}_{pj}$ or Y_{pj}, depending on whether or not the measured biomarker levels are subject to measurement error. Under the aforementioned assumptions, the probability density function of $\tilde{Y}_{pj}$, for a given value of $\boldsymbol{\theta} = (\boldsymbol{\beta}', \boldsymbol{\gamma}')'$, can be expressed as

$$f_j(\tilde{y}_{p_j}|\mathbf{x}_j,\boldsymbol{\theta})=\int\cdots\int c_j f\left(c_j\tilde{y}_{p_j}-\sum_{i=2}^{c_j}\tilde{y}_{ij}|\mathbf{x}_{1j},\boldsymbol{\theta}\right)\prod_{i=2}^{c_j}f(\tilde{y}_{ij}|\mathbf{x}_{ij},\boldsymbol{\theta})d\tilde{\boldsymbol{y}}_{(-1)j}, \quad (1)$$

where $\tilde{\boldsymbol{y}}_{(-1)j} = (\tilde{y}_{2j}, \ldots, \tilde{y}_{c_jj})$ and $\mathbf{x}_j = (\mathbf{x}_{1j}, \ldots, \mathbf{x}_{c_jj})$. In the presence of measurement error, it is common to assume that the conditional distribution of the observed measurement, given the true level, is known; cf., [33]. Let $f_\varepsilon(\cdot|\tilde{Y}_{pj})$ denote the conditional probability density function of Y_{pj}, given $\tilde{Y}_{pj}$. Thus, the density of Y_{pj} can be expressed as

$$g_j(y_{p_j}|\mathbf{x}_j,\boldsymbol{\theta})=\int\cdots\int f_\varepsilon\left(y_{p_j}|\tilde{y}_{p_j}\right)\prod_{i=1}^{c_j}f(\tilde{y}_{ij}|\mathbf{x}_{ij},\boldsymbol{\theta})d\tilde{\boldsymbol{y}}_j, \quad (2)$$

where $\tilde{\boldsymbol{y}}_j = (\tilde{y}_{1j}, \ldots, \tilde{y}_{c_jj})'$ and $\tilde{y}_{p_j}=c_j^{-1}\sum_{i=1}^{c_j}\tilde{y}_{ij}$. Note, under most of the common parametric models for biomarker data, expressions for $f_j(\cdot|\mathbf{x}_j, \boldsymbol{\theta})$ and $g_j(\cdot|\mathbf{x}_j, \boldsymbol{\theta})$ do not exist in closed form, but they can be evaluated through the use of numerical integration.

For ease of exposition, herein the proposed method is presented under the assumption that the observed pooled assessments are subject to measurement error. Although, it is worthwhile to point out that this approach is still applicable when the true biomarker levels are observed, as is demonstrated in Sections 4 and 5. Under the aforementioned modeling assumptions, the log of the observed data likelihood is given by

$$l(\boldsymbol{\theta}|\boldsymbol{Y}_p,\mathbf{x})=\sum_{j=1}^{J}\log\{g_j(Y_{p_j}|\mathbf{x}_j,\boldsymbol{\theta})\}, \quad (3)$$

where $\boldsymbol{Y}_p = (Y_{p1}, \ldots, Y_{pJ})'$ and $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_J)$. By maximizing (3) one can obtain the MLE of $\boldsymbol{\theta}_0$, which is denoted by $\hat{\boldsymbol{\theta}}$; i.e., $\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}\in\Theta} l(\boldsymbol{\theta}|\boldsymbol{Y}_p, \mathbf{x})$. If $g_j(\cdot|\mathbf{x}_j, \boldsymbol{\theta})$ could be expressed in closed form, obtaining the MLE of $\boldsymbol{\theta}_0$ would be relatively straightforward. Alternatively, if $g_j(\cdot|\mathbf{x}_j, \boldsymbol{\theta})$ does not exist in closed form, numerical integration techniques (e.g., adaptive Gaussian quadrature) could be used to evaluate $g_j(\cdot|\mathbf{x}_j, \boldsymbol{\theta})$ at Y_{pj}, and thus facilitate the maximization of (3). However, it is well known [34] that the computational burden associated with implementing these numerical techniques rapidly increases with the dimension of the integral, making this approach infeasible for $c_j > 2$; for further discussion see [30]. Further, numerical integration techniques, like the adaptive Gaussian quadrature, may perform poorly for peaked-integrand distribution functions [35]. Consequently, it is not recommended that these techniques be used to facilitate the maximization of (3).

To overcome the drawbacks of implementing numerical integration methods, such as adaptive Gaussian quadrature, the proposed approach uses a Monte Carlo technique to approximate the value of $g_j(\cdot|\mathbf{x}_j, \boldsymbol{\theta})$ when evaluated at Y_{pj}. Note, the value of $g_j(Y_{pj}|\mathbf{x}_j, \boldsymbol{\theta})$ can be viewed as the expected value of $f_\varepsilon(Y_{pj}|\tilde{Y}_{pj})$ with respect to $\tilde{Y}_j = (\tilde{Y}_{1j}, \ldots, \tilde{Y}_{c_jj})'$; i.e.

$$g_j(Y_{p_j}|\mathbf{x}_j,\boldsymbol{\theta})=E\{f_\varepsilon(Y_{p_j}|\tilde{Y}_{p_j})\}=\int\cdots\int f_\varepsilon(Y_{p_j}|\tilde{y}_{p_j})h_j(\tilde{\boldsymbol{y}}_j|\mathbf{x}_j,\boldsymbol{\theta})d\tilde{\boldsymbol{y}}_j, \tag{4}$$

where $h_j(\tilde{\boldsymbol{y}}_j|\mathbf{x}_j,\boldsymbol{\theta})=\prod_{i=1}^{c_j} f(\tilde{y}_{ij}|\mathbf{x}_{ij},\boldsymbol{\theta})$ is the joint density of $\tilde{\boldsymbol{Y}}_j$. Let $\tilde{\boldsymbol{Y}}_j^1,\ldots,\tilde{\boldsymbol{Y}}_j^M$ be a random sample from $h_j(\cdot|\mathbf{x}_j, \boldsymbol{\theta})$, where $\tilde{\boldsymbol{Y}}_j^m=(\tilde{Y}_{1j}^m,\ldots,\tilde{Y}_{c_jj}^m)'$, and define $\tilde{Y}_{p_j}^m=c_j^{-1}\sum_{i=1}^{c_j}\tilde{Y}_{ij}^m$. A Monte Carlo estimate of $g_j(Y_{pj}|\mathbf{x}_j, \boldsymbol{\theta})$ can then be obtained as

$$g_j^M(Y_{p_j}|\mathbf{x}_j,\boldsymbol{\theta})=\frac{1}{M}\sum_{m=1}^{M} f_\varepsilon(Y_{p_j}|\tilde{Y}_{p_j}^m). \tag{5}$$

Consequently, a Monte Carlo approximation of the log-likelihood, when evaluated at a specific value of $\boldsymbol{\theta}$, can be obtained as

$$l_M(\boldsymbol{\theta}|\mathbf{Y}_p,\mathbf{x})=\sum_{j=1}^{J}\log\{g_j^M(Y_{p_j}|\mathbf{x}_j,\boldsymbol{\theta})\}. \tag{6}$$

By numerically maximizing (6), one can obtain the Monte Carlo Maximum Likelihood Estimator (MCMLE) of $\boldsymbol{\theta}_0$, which is denoted by $\hat{\boldsymbol{\theta}}_M$; i.e., $\hat{\boldsymbol{\theta}}_M = \arg\max_{\boldsymbol{\theta}\in\Theta} l_M(\boldsymbol{\theta}|\mathbf{Y}_p, \mathbf{x})$. Further, under mild regularity conditions, see Section 3, it can be shown that $\hat{\boldsymbol{\theta}}_M \xrightarrow{p} \hat{\boldsymbol{\theta}}$, as $M \to \infty$.

Regretfully, for the purposes of maximizing (6), when (5) is being used to approximate $g_j(Y_{pj}|\mathbf{x}_j, \boldsymbol{\theta})$, numerical optimization algorithms can often be unreliable; i.e., these algorithms have the propensity to converge before reaching $\hat{\boldsymbol{\theta}}_M$. This is a byproduct of the fact that these algorithms require that the Monte Carlo log-likelihood be evaluated at multiple values of $\boldsymbol{\theta}$, and at each value of $\boldsymbol{\theta}$ a new random sample from $h_j(\tilde{\boldsymbol{Y}}_j|\mathbf{x}_j, \boldsymbol{\theta})$, for j = 1, …, J, has to be taken in order to evaluate (6). Due to the inherent variability in each of the random samples, this process results in a coarse (non-smooth) objective function which is difficult to numerically optimize; for further discussion see [36]. To circumvent this issue, [36] suggests that a single random sample, for each j, be drawn from a predetermined importance distribution $h_j^*(\cdot|\mathbf{x}_j, \boldsymbol{\theta}^*)$, where $\boldsymbol{\theta}^*$ is a known set of parameters. This random sample would then be used to evaluate (6) for each value of $\boldsymbol{\theta}$, thus insuring that the Monte Carlo log-likelihood is a smooth function and that optimization algorithms can be used reliably to obtain $\hat{\boldsymbol{\theta}}_M$.

Proceeding in this fashion, let $\tilde{\boldsymbol{Y}}_j^1, \ldots, \tilde{\boldsymbol{Y}}_j^M$ be a random sample from $h_j^*(\cdot|\mathbf{x}_j, \boldsymbol{\theta}^*)$, and based on this sample a Monte Carlo estimate of $g_j(Y_{pj}|\mathbf{x}_j, \boldsymbol{\theta})$ can be obtained as

$$g_j^M(Y_{p_j}|\mathbf{x}_j, \boldsymbol{\theta}, \boldsymbol{\theta}^*) = \frac{1}{M}\sum_{m=1}^{M} f_\varepsilon(Y_{p_j}|\tilde{Y}_{p_j}^m) w_j(\tilde{\boldsymbol{Y}}_j^m, \mathbf{x}_j, \boldsymbol{\theta}, \boldsymbol{\theta}^*) \qquad (7)$$

where $w_j(\tilde{\boldsymbol{Y}}_j^m, \mathbf{x}_j, \boldsymbol{\theta}, \boldsymbol{\theta}^*) = h_j(\tilde{\boldsymbol{Y}}_j^m|\mathbf{x}_j, \boldsymbol{\theta})/h_j^*(\tilde{\boldsymbol{Y}}_j^m|\mathbf{x}_j, \boldsymbol{\theta}^*)$. The corresponding Monte Carlo log-likelihood is given by

$$l_M(\boldsymbol{\theta}|\boldsymbol{Y}_p, \mathbf{x}, \boldsymbol{\theta}^*) = \sum_{j=1}^{J} \log\{g_j^M(Y_{p_j}|\mathbf{x}_j, \boldsymbol{\theta}, \boldsymbol{\theta}^*)\}, \qquad (8)$$

and the MCMLE of $\boldsymbol{\theta}_0$ can be obtained as $\hat{\boldsymbol{\theta}}_M = \arg\max_{\boldsymbol{\theta}\in\Theta} l_M(\boldsymbol{\theta}|\mathbf{Y}_p, \mathbf{x}, \boldsymbol{\theta}^*)$. Note, this process is identical to the approach described above, with the exception of the distributions from which the Monte Carlo samples were drawn.

Theoretically, there are very few restrictions on the choice of the importance distribution $h_j^*(\cdot|\mathbf{x}_j, \boldsymbol{\theta}^*)$, but its specification can dramatically impact the computational efficiency of the proposed approach. In general, the computational burden associated with the proposed approach is due to the size of the Monte Carlo sample being drawn to construct the estimator in (7). That is, obtaining the MCMLE can be computationally inefficient if M is too large, and alternatively imprecise if M is to small. In Section 3, guidance is provided on how to choose the value of M in order to insure that a specified level of precision is attained, under a specific importance distribution. Note, the Monte Carlo sample size required to attain the specified level of precision is inherently tied to the choice of the importance distribution; i.e., a well specified importance distribution results in a smaller value of M, and vice versa. In general, $h_j^*(\cdot|\mathbf{x}_j, \boldsymbol{\theta}^*)$ should have the same support as $h_j(\cdot|\mathbf{x}_j, \boldsymbol{\theta}_0)$, and the two densities should be similar in shape and center. Further, $h_j^*(\cdot|\mathbf{x}_j, \boldsymbol{\theta}^*)$ should be easy to sample from. In Sections 4 and 5, the importance distributions were selected according to the strategies considered in [37] and [30]; for further discussion see Web Appendix A. Note, under specific types of measurement error (e.g., additive) other more computationally efficient estimators akin to (7) can be derived; for further details see Web Appendix A.2.

3. Asymptotic properties

It is important to note that the Monte Carlo log-likelihood $l_M(\boldsymbol{\theta}|\mathbf{Y}_p, \mathbf{x}, \boldsymbol{\theta}^*)$ is evaluated using a random sample from the importance distribution and, as such, different random samples will yield different values of $\hat{\boldsymbol{\theta}}_M$. Therefore, $\hat{\boldsymbol{\theta}}_M$ is a non-deterministic approximation of $\hat{\boldsymbol{\theta}}$ whose variability is referred to as Monte Carlo error [38]. As was previously mentioned, the Monte Carlo sample size M plays a significant role with respect to determining the precision of $\hat{\boldsymbol{\theta}}_M$; i.e., the Monte Carlo error decreases as M is increased, and vice versa. By quantifying the Monte Carlo error, one can determine a value of M which will ensure that the approximation of $\hat{\boldsymbol{\theta}}$ attains a specified level of precision. To this end, the asymptotic properties of $\hat{\boldsymbol{\theta}}_M$, as $M \to \infty$, are presented and are further used to develop a method of identifying M which controls the Monte Carlo error at a specified level.

Throughout the remainder of this article, it is assumed that $\boldsymbol{\theta}$ consists of k components and that the parameter space $\boldsymbol{\Theta}$ is a compact subset of $\mathbb{R}^k$. From Kolmogorov's Strong Law of Large Numbers (SLLN), under standard regularity conditions, it is easy to establish that $\forall\ \boldsymbol{\theta} \in \boldsymbol{\Theta}$, the Monte Carlo log-likelihood function (8) converges, as $M \to \infty$, to the true log-likelihood function with probability 1 (w.p.1). However, this point-wise convergence does not guarantee the consistency of $\hat{\boldsymbol{\theta}}_M$; i.e., it does not establish that $\hat{\boldsymbol{\theta}}_M \xrightarrow{p} \hat{\boldsymbol{\theta}}$. A sufficient condition under which consistency can be established is the uniform convergence of (8), i.e.

$$\lim_{M\to\infty}\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}|l_M(\boldsymbol{\theta}|\boldsymbol{Y}_p,\mathbf{x},\boldsymbol{\theta}^*)-l(\boldsymbol{\theta}|\boldsymbol{Y}_p,\mathbf{x})|=0,\ \text{w. p. 1},$$

and Theorem 1 provides this result.

Theorem 1

Under regularity conditions i)–iii) provided in Web Appendix B, *it can be shown that*

$$\lim_{M\to\infty}\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}|l_M(\boldsymbol{\theta}|\boldsymbol{Y}_p,\mathbf{x},\boldsymbol{\theta}^*)-l(\boldsymbol{\theta}|\boldsymbol{Y}_p,\mathbf{x})|=0,\ \text{w. p.1.}$$

For a proof of Theorem 1 see [39]. Given the uniform convergence established in Theorem 1, the following result provides the consistency of $\hat{\boldsymbol{\theta}}_M$.

Corollary 1—*Let* $\hat{\boldsymbol{\theta}}$ *denote the unique element* $\boldsymbol{\Theta}$, *such that* $\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} l(\boldsymbol{\theta}|\mathbf{Y}_p, \mathbf{x})$, *and let* $\{\hat{\boldsymbol{\theta}}_M\}_M$ *be a sequence of maximizers of* $\{l_M(\boldsymbol{\theta}|\mathbf{Y}_p, \mathbf{x}, \boldsymbol{\theta}^*)\}_M$, *then* $\hat{\boldsymbol{\theta}}_M \xrightarrow{p} \hat{\boldsymbol{\theta}}$, *as* $M \to \infty$.

For a proof of Corollary 1 see [39]. Given the consistency of $\hat{\boldsymbol{\theta}}_M$, the asymptotic normality is established.

Theorem 2

Under regularity conditions iv)–viii) provided in Web Appendix B, *as $M \to \infty$, then*

$$\begin{array}{rcl}\sqrt{M}\nabla l_M(\hat{\boldsymbol{\theta}}|\,\boldsymbol{Y}_p,\mathbf{x},\boldsymbol{\theta}^*) & \xrightarrow{d} & N(\mathbf{0},\mathbf{A}),\\ -\nabla^2 l_M(\hat{\boldsymbol{\theta}}_M|\,\boldsymbol{Y}_p,\mathbf{x},\boldsymbol{\theta}^*) & \xrightarrow{p} & \mathbf{B},\\ \sqrt{M}(\hat{\boldsymbol{\theta}}_M-\hat{\boldsymbol{\theta}}|\boldsymbol{\theta}^*) & \xrightarrow{d} & N(\mathbf{0},\sum),\end{array}$$

where $\Sigma = \mathbf{B}^{-1}\mathbf{A}\mathbf{B}^{-1}$ *and* $\mathbf{B} = -\nabla^2 l(\hat{\boldsymbol{\theta}}|\mathbf{Y}_p, \mathbf{x})$.

The proof of Theorem 2 is classical [40] and a sketch is provided in Web Appendix C. Further, it is worthwhile to point out that it relatively easy to show that all of the parametric models considered herein adhere to the regularity conditions provided in Web Appendix C.

Establishing Theorem 2 provides two primary benefits; i.e., a means to identify the Monte Carlo sample size that maintains a specified level of precision and it allows one to perform typical large sample inference. Using the covariance matrix Σ, an asymptotic approximation of the Monte Carlo error associated with estimating $\hat{\boldsymbol{\theta}}$ can be obtained. Consequently, a natural strategy for choosing the Monte Carlo sample size would be to specify M so that the Monte Carlo error is bounded by a predetermined value, say d^2; i.e., choose M such that $\sigma^2_{\max}M^{-1} \le d^2$, where $\sigma^2_{\max}$ is the maximum diagonal element of Σ. Regretfully, in general a closed form expression for Σ does not exist, but it can be estimated. Under the regularity condition provided in Web Appendix B, a consistent estimator of $\mathbf{B}$ is given by $\hat{\mathbf{B}}(\hat{\boldsymbol{\theta}}_M|\boldsymbol{\theta}^*) = -\nabla^2 l_M(\hat{\boldsymbol{\theta}}_M|\mathbf{Y}_p, \mathbf{x}, \boldsymbol{\theta}^*)$, and in Web Appendix C an estimator of $\mathbf{A}$, which is denoted by $\hat{\mathbf{A}}(\hat{\boldsymbol{\theta}}_M|\boldsymbol{\theta}^*)$, is developed. Thus, an estimator of the covariance matrix Σ is given by

$$\hat{\sum}(\hat{\boldsymbol{\theta}}_M|\boldsymbol{\theta}^*)=\hat{\mathbf{B}}(\hat{\boldsymbol{\theta}}_M|\boldsymbol{\theta}^*)^{-1}\hat{\mathbf{A}}(\hat{\boldsymbol{\theta}}_M|\boldsymbol{\theta}^*)\hat{\mathbf{B}}(\hat{\boldsymbol{\theta}}_M|\boldsymbol{\theta}^*)^{-1}. \tag{9}$$

Using this estimator, M can be chosen such that $\hat{\sigma}^2_{\max}M^{-1} \le d^2$, where $\hat{\sigma}^2_{\max}$ is the maximum diagonal element of $\hat{\Sigma}(\hat{\boldsymbol{\theta}}_M|\boldsymbol{\theta}^*)$. Further, $\hat{\mathbf{B}}(\hat{\boldsymbol{\theta}}_M|\boldsymbol{\theta}^*)^{-1}$ is a consistent estimator of the observed Fisher information matrix; i.e., $\hat{\mathbf{B}}(\hat{\boldsymbol{\theta}}_M|\boldsymbol{\theta}^*)^{-1} \xrightarrow{p} \mathbf{B}^{-1}$, where $\mathbf{B} = -\nabla^2 l(\hat{\boldsymbol{\theta}}|\mathbf{Y}_p, \mathbf{x})$. Using this estimator, one can conduct typical Wald type inference.

One will note that the process of estimating Σ relies on obtaining $\hat{\boldsymbol{\theta}}_M$, which in turn depends on M. Consequently, for determining the appropriate Monte Carlo sample size, it is suggested that the following approach be implemented to ensure that a specified level of precision is attained.

Step 1: Choose a value of d^2, the initial Monte Carlo sample size M_0, and $\boldsymbol{\theta}^*$.

Step 2: Find the MCMLE, $\hat{\boldsymbol{\theta}}_{M_0}$, and calculate $\hat{\Sigma}(\hat{\boldsymbol{\theta}}_{M_0}|\boldsymbol{\theta}^*)$.

Step 3: Based on $\hat{\sigma}^2_{\max}$, the maximum diagonal element of $\hat{\Sigma}(\hat{\boldsymbol{\theta}}_{M_0}|\boldsymbol{\theta}^*)$:

a. Accept $\hat{\boldsymbol{\theta}}_{M_0}$ as the final estimate of $\hat{\boldsymbol{\theta}}$, if $\hat{\sigma}^2_{\max} M_0^{-1} \leq d^2$.

b. Proceed to Step 4, if $\hat{\sigma}^2_{\max} M_0^{-1} > d^2$.

Step 4: Reselect $\boldsymbol{\theta}^*$ based on $\hat{\boldsymbol{\theta}}_{M_0}$ and recompute $\hat{\Sigma}(\hat{\boldsymbol{\theta}}_{M_0} | \boldsymbol{\theta}^*)$.

Step 5: Choose M such that $\hat{\sigma}^2_{\max} M^{-1} \leq d^2$ and find $\hat{\boldsymbol{\theta}}_M$, the MCMLE of $\hat{\boldsymbol{\theta}}$.

Note, depending on the goals of the study, one could use this procedure either to control the precision of the whole parameter vector, or a subset of interest.

4. Simulation study

A simulation study was conducted in order to assess the finite sample performance of the proposed methodology. In this study, the following models for the biomarker levels of the individuals were considered:

$$\mathrm{M1}{:}\tilde{Y}_{ij}|\mathbf{x}_{ij} \sim \mathrm{N}(\mu_{ij}, \sigma^2), \mu_{ij} = \mathbf{x}_{ij}^{'}\boldsymbol{\beta}, \text{and } \boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)^{'} = (4, 1, 2)^{'};$$
$$\mathrm{M2}{:}\log(\tilde{Y}_{ij})|\mathbf{x}_{ij} \sim \mathrm{N}\{\mu_{ij}, \sigma^2\}, \mu_{ij} = \mathbf{x}_{ij}^{'}\boldsymbol{\beta}, \text{and } \boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)^{'} = (-1, 1, 2)^{'};$$
$$\mathrm{M3}{:}\tilde{Y}_{ij}|\mathbf{x}_{ij} \sim \mathrm{ST}(\mu_{ij}, \nu), \mu_{ij} = \mathbf{x}_{ij}^{'}\boldsymbol{\beta}, \text{and } \boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)^{'} = (4, 1, 2)^{'};$$
$$\mathrm{M}_4{:}\tilde{Y}_{ij}|\mathbf{x}_{ij} \sim \mathrm{Gamma}\{1/\sigma^2, (\sigma^2\mu_{ij})^{-1}\}, \mu_{ij} = \eta_1^{-1}(\mathbf{x}_{ij}^{'}\boldsymbol{\beta}), \text{and } \boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)^{'} = (-1, 1, 2)^{'};$$
$$\mathrm{M}_5{:}\tilde{Y}_{ij}|\mathbf{x}_{ij} \sim \mathrm{Gamma}\{1/\sigma^2, (\sigma^2\mu_{ij})^{-1}\}, \mu_{ij} = \eta_2^{-1}(\mathbf{x}_{ij}^{'}\boldsymbol{\beta}), \text{and } \boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)^{'} = (0.5, 0.1, 0.2)^{'};$$

for $i = 1, \ldots, c_j$ and $j = 1, \ldots, J$, where $\eta_1(\cdot)$ and $\eta_2(\cdot)$ are the inverse and log links, respectively, $\mathbf{x}_{ij} = (1, x_{ij1}, x_{ij2})'$, $\sigma = 0.5$, and $\nu = 4$. Note, ST(μ, ν) denotes the shifted t-distribution, which has mean μ and degrees of freedom ν. The two covariates were independently sampled: one, x_{ij1}, from a standard normal distribution (which was supposed to emulate a standardized age effect) and the other, x_{ij2}, from a Bernoulli distribution having success probability 0.5 (which was supposed to correspond to a gender effect). For the purposes of this study, a common pool size c was specified, where $c \in \{1, \ldots, 5\}$; i.e., $c_j = c$, for $j = 1, \ldots, J$, where $J \in \{50, 100, 200, 500\}$. These choices were made to investigate the performance of the proposed approach across a broad spectrum of characteristics which could be encountered when modeling pooled biomarker data. In particular, the simulation configurations consider models both within (M1, M4, and M5) and outside (M2 and M3) of the class of generalized linear models which are commonly used to analyze biomarker data. Further, the different combinations of (c, J) provide an assessment of the impact of the pool and sample size on estimation and inference.

In both the binary group testing [31] and the pooled biomarker regression [30] literature it has been shown that pool composition, with respect to the covariates, has the potential to influence estimation efficiency. That is, randomly assigning subjects to pools, so that pools are heterogeneous with respect to covariate composition, can result in a loss in efficiency, while homogeneous pooling strategies (i.e., strategies which specify the pooling of subjects with similar covariates) maintain a high level of efficiency. For this reason, herein homogeneous pooling was used to assign subjects to pools. Once pool assignment was complete, the biomarker levels of the pools were determined as $\tilde{Y}_{p_j} = c^{-1}\sum_{i=1}^{c}\tilde{Y}_{ij}$, for $j =$

1, …, J. To investigate the effect of measurement error, an error contaminated measurement of $\tilde{Y}_{pj}$ was obtained as $Y_{pj}\,|\,\tilde{Y}_{pj} \sim N(\tilde{Y}_{pj}, \tau_2)$, for j = 1, …, J and $\tau \in \{0.05, 0.10, 0.20\}$. Combining these simulated values two separate pooled data sets were created: $\{(\tilde{Y}_{pj}, \mathbf{x}_j), j =$ 1, …, $J\}$ and $\{(Y_{pj}, \mathbf{x}_j), j =$ 1, …, $J\}$. This process was repeated 500 times for each model and configuration of (c, J, τ), resulting in 200,000 simulated data sets.

The methodology proposed in Section 2 was then implemented to analyze each of the aforementioned data sets. Throughout, the importance distributions for models M1–M5 were selected according to the strategies outlined in [37] and [30]; see Web Appendix A for their explicit forms and further discussion. To implement the model fitting approach described in Section (3), an initial Monte Carlo sample size of M_0 = 2000 was used and the tolerance was specified to be d^2 = 0.01. Note: to model the non-error contaminated data, minor alterations to the proposed approach are necessary. In particular, the density of $\tilde{Y}_{pj}$, which is given in (1), was approximated by

$$f_j^M(\tilde{Y}_{p_j}|\mathbf{x}_j,\boldsymbol{\theta},\boldsymbol{\theta}^*)=\frac{1}{M}\sum_{m=1}^{M}c_j f\left(c_j\tilde{Y}_{p_j}-\sum_{i=2}^{c_j}\tilde{Y}_{ij}^m|\mathbf{x}_{1j},\boldsymbol{\theta}\right)\frac{h_j(\tilde{\boldsymbol{Y}}_{(-1)j}^m|\mathbf{x}_j,\boldsymbol{\theta})}{h_j^*(\tilde{\boldsymbol{Y}}_{(-1)j}^m|\mathbf{x}_j,\boldsymbol{\theta}^*)},$$

where $\tilde{\boldsymbol{Y}}_{(-1)j}^m=(\tilde{Y}_{2j}^m,\ldots,\tilde{Y}_{c_jj}^m)'$, for m = 1, …, M, is a random sample from the importance distribution $h_j^*(\cdot|\mathbf{x}_j,\boldsymbol{\theta}^*)$. Replacing $g_j^M(Y_{p_j}|\mathbf{x}_j,\boldsymbol{\theta},\boldsymbol{\theta}^*)$ in (8) by $f_j^M(\tilde{Y}_{p_j}|\mathbf{x}_j,\boldsymbol{\theta},\boldsymbol{\theta}^*)$ and maximizing with respect to $\boldsymbol{\theta}$ results in obtaining an MCMLE of $\boldsymbol{\theta}_0$ when the pooled assessments are non-error laden. To complete model fitting for the error contaminated data it was assumed, as is common in the literature [22, 27], that the distribution of the measurement error was known. In practice this assumption may not be reasonable, but the true form of this distribution could be replaced by an estimate, which can be obtained through standard techniques; e.g., see [33]. Further, this approach is implemented in Section 5. Note, since additive measurement error is assumed the alternate formulation of (4), which is presented in Web Appendix A.2, is implemented herein.

In order to assess the performance of the proposed approach, it is first noted that when c = 1 and the $\tilde{Y}_{pj}$ are observed (i.e., the individual level data is observed without measurement error), standard regression techniques are applicable, and were implemented. This was done in order to provide a baseline by which comparisons could be made. Further, under the considered simulation configurations for model M1 it is possible to obtain an analytical expression for the MLE of $\boldsymbol{\theta}_0$ based on either $\{(\tilde{Y}_{pj}, \mathbf{x}_j), j =$ 1, …, $J\}$ or $\{(Y_{pj}, \mathbf{x}_j), j =$ 1, …, $J\}$, for c = 1, …, 5; see Web Appendix D for further details. Consequently, both the proposed approach and the aforementioned analytical techniques were used to analyze the simulated data created under model M1. Comparisons between the results obtained from these two techniques allows one to assess the error that is introduced into the analysis by stochastically estimating the MLE through the proposed methodology. Lastly, the MCEM algorithm proposed in [30] was used to analyze the non-error contaminated data under model M2; i.e., $\{(\tilde{Y}_{pj}, \mathbf{x}_j), j =$ 1, …, $J\}$. This competing approach was specifically designed for analyzing data of this form and this comparison allows one to assess the performance of the proposed methodology with respect to an existing technique.

Table 1 summarizes the estimates of the regression coefficients (i.e., β) that were obtained from analyzing the non-error contaminated pooled observations via the proposed approach and the alternate techniques, for all considered models and values of c, when $J = 100$. In particular, Table 1 summarizes the empirical bias and the sample standard deviation of the 500 point estimates of the regression coefficients stratified by model and pool size. The average of the 500 estimated standard errors, and the estimated coverage probability associated with 95% Wald confidence intervals are also included. Web Tables 1 and 2 in the Web Appendix provide the analogous results for $J = 50$ and 200, respectively. From these results, one will first notice that the point estimates obtained by the proposed approach exhibit little if any evidence of bias. Secondly, the sample standard deviation and the average standard error of the 500 estimates are predominantly in agreement. Further, the estimated coverage probabilities are all at their nominal level. These finding suggest that the approach proposed in Section 3 for estimating the asymptotic variance performs well for finite samples, and could subsequently be used to reliably conduct Wald-type inference.

To assess the effect that pooling has on parameter estimation and inference, one may compare the parameter estimates obtained from the individual level data (i.e., when $c = 1$) to those obtained from the pooled data (i.e., when $c > 1$). These comparisons reveal two striking features of the proposed methodology. First, the estimates of the regression coefficients based on the pooled data are more efficient (i.e., have a smaller sample standard deviation and average standard error) than the estimates obtained from analyzing the individual level data. Moreover, the efficiency of the estimators tends to increase with the pool size c. This finding suggests that the process of analyzing biomarker assessments made on pools, rather than individuals, will result in more precise estimation and inference, when a fixed number of assessments J is mandated. Secondly, this comparison also suggests that the error that is introduced by approximating the observed data likelihood through the expression in (8), is appropriately controlled by the approach presented in Section 3. This assertion is reinforced when one considers the comparison between the estimates obtained from the proposed methodology and the analytical form of the MLE under model M1. In particular, the summary measures of the estimated regression coefficients obtained from these two techniques are practically identical. This finding provides evidence that the approach developed in Section 3 can be used to appropriately control the precision of the MCMLE.

Table 1 also provides a summary of the estimated regression coefficients obtained from the MCEM algorithm proposed by [30], for model M2. In comparing these results to the those obtained from the proposed methodology one will note that the two procedures are practically identical in terms of estimation and inference. Though similar in terms of estimation and inference, an advantage of the proposed methodology over that of the MCEM algorithm arises in the computational time required to complete model fitting. Figure 1 provides the average model fitting time for both the proposed approach and the MCEM algorithm, for all considered combinations of (c, J) under model M2. From these results one will note that the proposed approach is able to complete model fitting roughly 5 to 8 times faster than the MCEM algorithm, on average. Moreover, it appears that the average model fitting time for the MCEM algorithm increases more rapidly with both the sample size (J) and pool size (c). Note, the aforementioned comparisons between the proposed approach and the MCEM algorithm were implemented using code written solely in R, which used no advanced computing techniques (e.g., interfacing R with C++, parallel processing, etc.), and was run on a Optiplex 790 desktop running Windows 7 with an Intel i7-2600 3.40 GHz CPU and 16GB of RAM. Further, these comparisons are for data that are not subject to measurement error, as the MCEM algorithm was not designed to correct for the effect of error laden measurements.

Table 2 summarizes the estimates of the regression coefficients that were obtained from analyzing the error contaminated pooled observations via the proposed approach, for all considered models and values of c, when $J = 100$ and $\tau = 0.05$. Web Tables 3–10 in the Web Appendix provides the analogous results for the other considered values of J and τ. The results from this study reinforce all of the main findings discussed above; i.e., the proposed approach can be used to accurately and efficiently analyze pooled biomarker data, while correcting for the effects of measurement error. In summary, the results of this study highlight the three definitive advantages of the proposed methodology, when compared to other existing techniques: first, the proposed technique can account for data subject to measurement error; second, the methodology outlined in Section 2 can be implemented under many different parametric models to conduct the regression analysis of measurements taken on pools; and third, the proposed methodology is far less computationally burdensome when compared to existing techniques.

5. Data application

The Collaborative Perinatal Project (CPP) was a longitudinal study, conducted from 1957 to 1974, which was aimed at assessing multiple hypotheses regarding varying aspects of maternal and child health [41]. The data collected by this study constitutes an important resource for biomedical research in many areas of perinatology and pediatrics. In 2007 data from the CPP was used in a nested case-control study which examined whether circulating levels of chemokines are related to miscarriage risk; for further details see [42]. In particular, this study considered measuring cytokine levels, to include monocyte chemotactic protein-1 (MCP1), on stored serum samples from CPP participants who had experienced a miscarriage (cases) and those who had not (controls), where cases and controls were matched based on gestational age. This analysis focuses on the MCP1 measurements obtained in the aforementioned study and considers only the participants for which full covariate information was available, where the selected explanatory variables consist of age (standardized; denoted by x_1), race (1=Africa American/0=otherwise; denoted by x_2), and miscarriage status (1=yes/0=no; denoted by x_3).

This data set possess two unique features which are of particular interest; i.e., MCP1 measurements were taken on both individual and pooled specimens, in duplicate. Consequently, the data available from this study can be divided into two separate data sets: the pooled data (PD) which consists of 81 and 350 MCP1 measurements taken on individual and pooled (with $c_j = 2$ for all j) specimens, respectively, and the individual data (ID) which consists of 752 measurements taken on individual specimens. This feature allows one to asses the effect of pooling on parameter estimation, a characteristics that could not be examined if only pooled data was considered. Further, within the PD and ID the MCP1 measurements were taken in duplicate for most of the specimens (pooled and individual), which allowed for the detection and subsequently the estimation of the measurement error [33]. In particular, through differencing the replicated MCP1 measurements it was found that it was reasonable to assume that additive measurement error was present, and that the error terms followed a normal distribution with mean 0 and variance τ^2. From these differences, τ^2 was estimated to be 0.044^2 and 0.050^2 based on the ID and PD, respectively.

As with many cytokines, the MCP1 measurements are positive and seem to follow a right-skewed distribution. For this reason, three parametric models were considered for the MCP1 levels: a log-normal regression model and a gamma regression model under two different link functions, namely the log and inverse links. The linear predictor of the full model was specified to be

$$\eta=\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3+\beta_4x_1^2+\beta_5x_1x_2+\beta_6x_1x_3+\beta_7x_2x_3,$$

where $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_7)'$ are the regression coefficients. To identify the final model, best subsets regression, driven by the Akaike information criterion, was implemented to evaluate each of the $2^7 = 128$ submodels. In order to assess the effect of measurement error, this process, for each parametric model, was completed both acknowledging and ignoring measurement error for both data sets (i.e., ID and PD). Consequently, this analysis considers 12 different scenarios, each of which required the fitting of 128 models. For all of these analyses, the proposed methodology was implemented in the same fashion as was discussed in Section 4.

Table 3 provides the top 5 models (ranked according to their AIC values) within each scenario. From these results, it can be seen that under all considered parametric models, both for the ID and PD, the collection of predictor variables most frequently selected include age (x_1), miscarriage status (x_3), and the interaction between miscarriage status and race (x_2x_3). For this reason, the final model under all scenarios was chosen to possess these predictor variables. For comparative purposes, the MCEM algorithm proposed by [30] was also implemented to perform model fitting and evaluation for the PD under the log-normal regression model, assuming that the observed outcomes were non-error laden. The model estimates, and consequently the AIC values, obtained by both the MCEM algorithm and the proposed methodology were practically identical. The marked difference between these two competing techniques arose in the computational time required to complete best subsets regression. In particular, the MCEM algorithm required 74 hours to complete this process, while the proposed approach took only 7. Consequently, based on these findings it is conjectured that the proposed methodology would likely be preferred in practice due to its computational efficiency, especially for the purposes of completing model selection via automated search algorithms.

Table 4 provides the estimates of the regression coefficient along with their estimated standard errors and p-values for the final model, across the 12 different scenarios. These results warrant several comments, first one will note that ignoring measurement error in this application impacted both estimation and inference, for both the ID and PD; i.e., there were 4 (1) cases for the ID (PD) in which a regression coefficient was deemed to be significant (at the $\alpha = 0.05$ significance level) when measurement error was accounted for but were found to be insignificant when it was ignored. Secondly, the estimates obtained from the PD are in general agreement with the estimates resulting from the ID, across all of the considered configurations. Although, in some instances discrepancies were observed; e.g., under the gamma regression model with the inverse link, the estimates of the regression coefficient associated with the interaction term obtained from the ID and PD were 1.972 and 7.002, respectively.

One plausible explanation for these deviations involves the pooling strategy considered in the original study; i.e. cases were randomly pooled with cases and controls with controls. In order to investigate this assertion, a second pooled data (HPD) set was artificially constructed using the ID, where pools, of size 2, were formed homogeneously with respect to the participants covariate information and the MCP1 measurement for each pool was taken to be the average of the MCP1 measurements for the individuals of which it was comprised. Table 4 provides the parameter estimates obtained from the analysis of the HPD, and from these results one will note that the use of homogeneous pooling has practically resolved all of the aforementioned discrepancies; i.e., the regression parameter estimates based on ID and HPD are practically identical. Further, the regression parameter estimates obtained from the HPD were generally more accurate and efficient than those based on the PD which contains 80 more observations.

The aforementioned analysis illustrates the primary strengths of the proposed approach; i.e., this new methodology can be used to conduct the regression analysis of pooled data under a variety of parametric models, it can be used to directly account for the effects of measurement error, and it is computationally efficient. Further, this study also illustrates that in order to obtain the most accurate estimation and inference homogeneous pooling should be implemented in practice. Similar findings were reported in [30].

6. Discussion

In this work, a general framework for the regression analysis of assessments taken on pools has been developed. The proposed approach allows for the regression analysis under practically all parametric models, can be used to account for the effects of measurement error if present, and is computationally efficient. The asymptotic properties of the proposed technique have been established. Through simulation studies the proposed approach was shown to obtain precise estimates and accurate inference under several common parametric models, as well as to be superior, with respect to computational efficiency, to existing techniques. Further, the simulation study indicated that the regression analysis of pooled data can result in parameter estimates that are more efficient than those which are based on individual level data, if pools are formed homogeneously and a fixed number of assessments, J, are to be made. Similarly, from the data application one will observe that for a fixed population size, there is practically no loss in estimation efficiency when analyzing homogeneously pooled outcomes. In order to further disseminate this work, software, programmed in R, which implements the proposed methodology has been developed and is available upon request.

The primary focus of the work presented in this manuscript was placed on developing a general methodology for conducting the regression analysis of pooled assessments, based on an assumed parametric model for the latent individual level data. A topic for future research could involve developing techniques which can be used to evaluate the selected parametric model; e.g., goodness-of-fit tests. Further, similar to [29], future research in this area could focus on developing pooling algorithms which attempt to minimize the loss in estimation efficiency that is incurred due to pooling.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This research was partially supported by Grant R01 AI121351 from the National Institutes of Health.

References

1. Dorfman R. The detection of defective members of large populations. Annals of Mathematical Statistics. 1943; 14:436–440.
2. Cardoso M, Koerner K, Kubanek B. Mini-pool screening by nucleic acid testing for hepatitis B virus, hepatitis C virus, and HIV: Preliminary results. Transfusion. 1998; 38:905–907. [PubMed: 9767739]
3. Lewis J, Lockary V, Kobic S. Cost savings and increased efficiency using a stratified specimen pooling strategy for *Chlamydia trachomatis* and *Neisseria gonorrhoeae*. Sexually Transmitted Diseases. 2012; 39:46–48. [PubMed: 22183846]
4. Van T, Miller J, Warshauer D, Reisdorf E, Jerrigan D, Humes R, Shult P. Pooling nasopharyngeal/throat swab speciments to increase testing capacity for influenza viruses by PCR. Journal of Clinical Microbiology. 2012; 50:891–896. [PubMed: 22205820]
5. Stramer S, Notari E, Krysztof D, Dodd R. Hepatitis B virus testing by minipool nucleic acid testing: does it improve blood safety. Transfusion. 2013; 53:2525–2537. [PubMed: 23550838]
6. Schmidt M, Roth W, Meyer H, Seifried E, Hourfar M. Nucleic acid test screening of blood donors for orthopoxviruses can potentially prevent dispersion of viral agents in case of bioterrorism. Transfusion. 2005; 45:399–403. [PubMed: 15752158]
7. Remlinger K, Hughes O, Young S, Lam R. Statistical design of pools using optimal coverage and minimal collision. Technometrics. 2006; 48:133–143.
8. Gastwirth J. The efficiency of pooling in the detection of rare mutations. The American Journal of Human Genetics. 2000; 67:1036–4039. [PubMed: 10986050]
9. Caudill S. Use of pooled samples from the national health and nutrition examination survey. Statistics in medicine. 2012; 31:3269–3277. [PubMed: 22492247]
10. Thompson K. Estimation of the proportion of vectors in a natural population of insects. Biometrics. 1962; 18:568–578.
11. Walter S, Hildreth S, Beaty B. Estimation of infection rates in populations of organisms using pools of variable size. American Journal of Epidemiology. 1980; 112:124–128. [PubMed: 7395846]

12. Brookmeyer R. Analysis of multistage pooling studies of biological specimens for estimating disease incidence and prevalence. Biometrics. 1999; 55:608–612. [PubMed: 11318222]
13. Farrington C. Estimating prevalence by group testing using generalized linear models. Statistics in medicine. 1992; 11:1591–1597. [PubMed: 1359621]
14. 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]
15. Xie M. Regression analysis of group testing samples. Statistics in Medicine. 2001; 20:1957–1969. [PubMed: 11427952]
16. Chen P, Tebbs JM, Bilder CR. Group testing regression models with fixed and random effects. Biometrics. 2009; 65:1270–1278. [PubMed: 19210734]
17. Huang X, Tebbs JM. On latent-variable model misspecification in structural measurement error models for binary response. Biometrics. 2009; 65:710–718. [PubMed: 18945265]
18. Delaigle A, Meister A. Nonparametric regression analysis for group testing data. Journal of the American Statistical Association. 2011; 106:640–650.
19. Delaigle A, Hall P. Nonparametric regression with homogeneous group testing data. Annals of Statistics. 2012; 40:131–158.
20. Wang D, McMahan C, Gallagher C, Kulasekera KB. Semiparametric group testing regression models. Biometrika. 2014; 101:587–598.
21. Delaigle A, Hall P, Wishart J. New approaches to nonparametric and semiparametric regression for univariate and multivariate group testing data. Biometrika. 2014; 101:567–585.
22. Schisterman E, Faraggi D, Reiser B, Trevisan M. Statistical inference for the area under the receiver operating characteristic curve in the presence of random measurement error. Annals of Epidemiology. 2001; 154:174–179.
23. Faraggi D, Reiser B, Schisterman E. ROC curve analysis for biomarkers based on pooled assessments. Statistics in Medicine. 2003; 22:2515–2527. [PubMed: 12872306]
24. Liu A, Schisterman E. Comparison of diagnostic accuracy of biomarkers with pooled assessments. Biometrical Journal. 2003; 45:631–644.
25. Mumford S, Schisterman E, Vexler A, Liu A. Pooling biospecimens and limits of detection: Effects on ROC curve analysis. Biostatistics. 2006; 7:585–598. [PubMed: 16531470]
26. Bondell H, Liu A, Schisterman E. Statistical inference based on pooled data: a moment-based estimating equation approach. Journal of Applied Statistics. 2007; 34:129–140.
27. Vexler A, Schisterman E, Liu A. Estimation of ROC curves based on stably distributed biomarkers subject to measurement error and pooling mixtures. Statistics in Medicine. 2008; 27:280–296. [PubMed: 17721905]
28. Malinovsky Y, Albert P, Schisterman E. Pooling designs for outcomes under a Gaussian random effects model. Biometrics. 2012; 68:45–52. [PubMed: 21981372]
29. Ma C, Vexler A, Schisterman E, Tian L. Cost-efficient designs based on linearly associated biomarkers. Journal of Applied Statistics. 2011; 38:2739–2750.
30. Mitchell E, Lyles R, Manatunga A, Danaher M, Perkins N, Schisterman E. Regression for skewed biomarker outcomes subject to pooling. Biometrics. 2014; 70:202–211. [PubMed: 24521420]
31. Zhang B, Bilder CR, Tebbs JM. Group testing regression model estimation when case identification is a goal. Biometrical Journal. 2013; 55:173–189. [PubMed: 23401252]
32. McCullagh, P., Nelder, JA. Generalized linear models. CRC press; 1989.
33. Carroll, R., Ruppert, D., Stefanski, L., Crainiceanu, C. Measurement Error in Nonlinear Models: A Modern Perspective. 2. London: Chapmen & Hall; 2006.
34. Kuonen D. Numerical Integration in S-PLUS or R: A Survey. Journal of Statistical Software. 2003; 8:1–14.
35. Genz A, Kass R. Subregion-adaptive integration of functions having a dominant peak. Journal of Computational and Graphical Statistics. 1997; 6:92–111.
36. Robert, C., Casella, G. Monte Carlo statistical methods. New York: Springer; 1999.

37. Frigyik, B., Kapila, A., Gupta, M. UWEE Technical Report UWEETR-2010-0006. Department of Electrical Engineering, University of Washington; Seattle, WA: Introduction to the dirichlet distribution and related processes.

38. Booth J, Hobert J. Maximizing generalized linear mixed model likelihoods with an automated Monte Carlo EM algorithm. Journal of the Royal Statistical Society, Series B. 1999; 61:265–285.

39. Beskos A, Papaspiliopoulos O, Roberts G. Monte Carlo maximum likelihood estimation for discretely observed diffusion processes. The Annals of Statistics. 2009; 37:223–245.

40. Geyer C. On the convergence of Monte Carlo maximum likelihood calculations. Journal of the Royal Statistical Society, Series B. 1994; 56:261–274.

41. Hardy J. The collaborative perinatal project: Lessons and legacy. Annals of Epidemiology. 2003; 5:303–311.

42. Whitcomb B, Schisterman E, Klebanoff M, Baumgarten M, Rhoton-Vlasak A, Luo X, Chegini N. Circulating chemokine levels and miscarriage. American Journal of Epidemiology. 2007; 166:323–331. [PubMed: 17504778]

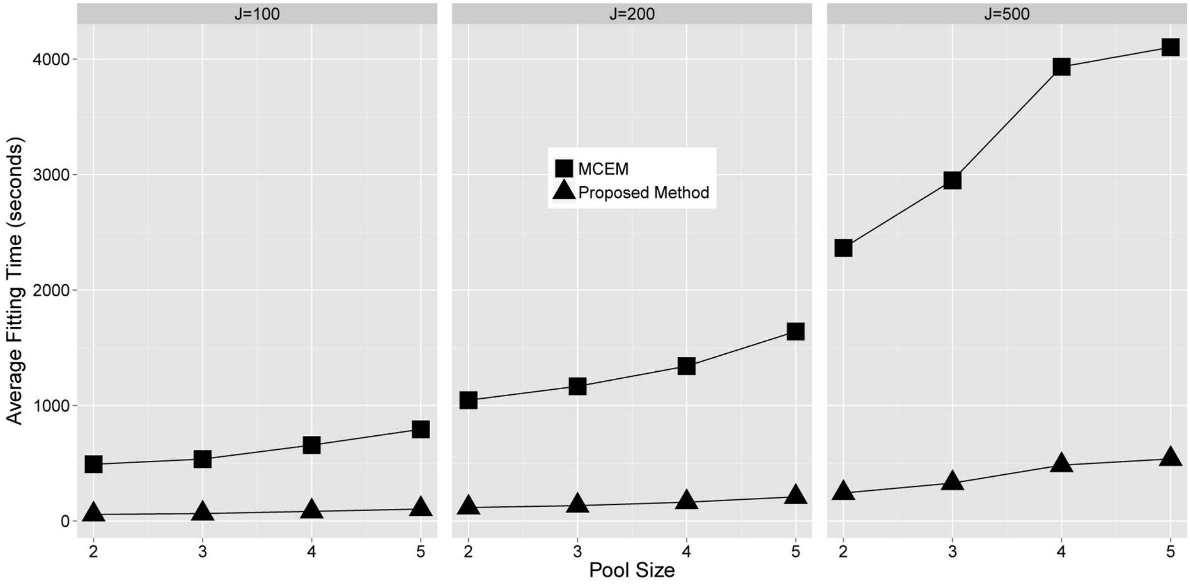

Figure 1.

Simulation study: Average model fitting times required by the proposed methodology and the MCEM algorithm developed in Mitchell et al. (2014) for data generated under model M2 which is not subject to measurement error. Presented results are stratified by pool size (*c*) and the number of pools (*J*).

Table 1

Presented results include the empirical bias (Bias) of the 500 estimated regression coefficients and there sample standard deviation (SD) obtained from analyzing data generated according to models M1–M5, for all considered pool sizes when $J = 100$ in the absence of measurement error. Also included are the average estimated standard errors (SE) and the empirical coverage probabilities (Cov) associated with 95% Wald confidence intervals. Three model fitting procedures were implemented, the proposed methodology (MCMLE), the analytical approach described in Section 4 (MLE), and the MCEM algorithm, with the latter two techniques only being applicable for models M1 and M2, respectively.

Model		Measure	$c = 1$	$c = 2$	$c = 3$	$c = 4$	$c = 5$
M1(MCMLE)	β_0	Bias(SD)	0.00(0.07)	0.00(0.05)	0.00(0.04)	0.00(0.04)	0.00(0.03)
		Cov(SE)	0.96(0.07)	0.94(0.05)	0.94(0.04)	0.95(0.03)	0.92(0.03)
	β_1	Bias(SD)	0.00(0.05)	0.00(0.03)	0.00(0.03)	0.00(0.03)	0.00(0.02)
		Cov(SE)	0.96(0.05)	0.95(0.04)	0.93(0.03)	0.94(0.02)	0.94(0.02)
	β_2	Bias(SD)	0.00(0.10)	0.00(0.07)	0.00(0.06)	0.00(0.05)	0.00(0.05)
		Cov(SE)	0.94(0.10)	0.94(0.07)	0.94(0.06)	0.94(0.05)	0.93(0.04)
M1(MLE)	β_0	Bias(SD)	0.00(0.07)	0.00(0.05)	0.00(0.04)	0.00(0.04)	0.00(0.03)
		Cov(SE)	0.96(0.07)	0.95(0.05)	0.94(0.04)	0.95(0.04)	0.92(0.03)
	β_1	Bias(SD)	0.00(0.05)	0.00(0.03)	0.00(0.03)	0.00(0.03)	0.00(0.02)
		Cov(SE)	0.96(0.05)	0.96(0.04)	0.94(0.03)	0.95(0.03)	0.94(0.02)
	β_2	Bias(SD)	0.00(0.10)	0.00(0.07)	0.00(0.06)	0.00(0.05)	0.00(0.05)
		Cov(SE)	0.94(0.10)	0.95(0.07)	0.94(0.06)	0.94(0.05)	0.93(0.04)
M2(MCMLE)	β_0	Bias(SD)	0.00(0.07)	0.00(0.05)	0.00(0.04)	0.00(0.04)	0.00(0.04)
		Cov(SE)	0.94(0.07)	0.95(0.05)	0.95(0.04)	0.94(0.04)	0.94(0.03)
	β_1	Bias(SD)	0.00(0.05)	0.00(0.04)	0.00(0.03)	0.00(0.03)	0.00(0.02)
		Cov(SE)	0.95(0.05)	0.94(0.04)	0.95(0.03)	0.94(0.03)	0.94(0.02)
	β_2	Bias(SD)	0.01(0.10)	0.00(0.08)	0.00(0.06)	0.00(0.05)	0.00(0.05)
		Cov(SE)	0.95(0.10)	0.93(0.07)	0.96(0.06)	0.97(0.05)	0.96(0.05)
M2(MCEM)	β_0	Bias(SD)	0.00(0.07)	0.00(0.05)	0.00(0.04)	0.00(0.04)	0.00(0.04)
		Cov(SE)	0.94(0.07)	0.95(0.05)	0.95(0.04)	0.93(0.04)	0.94(0.03)
	β_1	Bias(SD)	0.00(0.05)	0.00(0.04)	0.00(0.03)	0.00(0.03)	0.00(0.02)
		Cov(SE)	0.95(0.05)	0.94(0.04)	0.94(0.03)	0.94(0.03)	0.94(0.02)
	β_2	Bias(SD)	0.01(0.10)	0.00(0.08)	0.00(0.06)	0.00(0.05)	0.00(0.05)
		Cov(SE)	0.95(0.10)	0.92(0.07)	0.96(0.06)	0.97(0.05)	0.96(0.05)

Model		Measure	*c* = 1	*c* = 2	*c* = 3	*c* = 4	*c* = 5
M3(MCMLE)	β_0	Bias(SD)	0.00(0.16)	−0.01(0.12)	0.00(0.11)	0.00(0.10)	0.00(0.09)
		Cov(SE)	0.96(0.16)	0.95(0.13)	0.94(0.11)	0.95(0.09)	0.94(0.08)
	β_1	Bias(SD)	0.00(0.12)	0.00(0.09)	0.00(0.08)	0.00(0.06)	0.00(0.06)
		Cov(SE)	0.94(0.12)	0.94(0.09)	0.96(0.08)	0.95(0.07)	0.94(0.06)
	β_2	Bias(SD)	0.01(0.24)	0.01(0.18)	0.00(0.16)	0.00(0.14)	0.00(0.13)
		Cov(SE)	0.94(0.23)	0.93(0.18)	0.93(0.15)	0.95(0.13)	0.95(0.12)
M4(MCMLE)	β_0	Bias(SD)	0.00(0.07)	0.00(0.05)	0.00(0.04)	0.00(0.04)	0.00(0.03)
		Cov(SE)	0.96(0.07)	0.94(0.05)	0.94(0.04)	0.94(0.03)	0.94(0.03)
	β_1	Bias(SD)	0.00(0.05)	0.00(0.03)	0.00(0.03)	0.00(0.03)	0.00(0.02)
		Cov(SE)	0.96(0.05)	0.98(0.04)	0.93(0.03)	0.95(0.03)	0.93(0.02)
	β_2	Bias(SD)	0.00(0.10)	0.00(0.07)	0.00(0.06)	0.00(0.05)	0.00(0.04)
		Cov(SE)	0.95(0.10)	0.95(0.07)	0.95(0.06)	0.95(0.05)	0.95(0.04)
M5(MCMLE)	β_0	Bias(SD)	0.01(0.04)	0.00(0.03)	0.00(0.02)	0.00(0.02)	0.00(0.02)
		Cov(SE)	0.94(0.04)	0.94(0.02)	0.95(0.02)	0.92(0.02)	0.96(0.02)
	β_1	Bias(SD)	0.00(0.03)	0.00(0.02)	0.00(0.01)	0.00(0.01)	0.00(0.01)
		Cov(SE)	0.94(0.03)	0.94(0.02)	0.95(0.01)	0.96(0.01)	0.97(0.01)
	β_2	Bias(SD)	0.00(0.06)	0.00(0.04)	0.00(0.03)	0.00(0.03)	0.00(0.02)
		Cov(SE)	0.95(0.06)	0.95(0.04)	0.94(0.03)	0.93(0.03)	0.97(0.03)

Table 2

Simulation study: Presented results include the empirical bias (Bias) of the 500 estimated regression coefficients and there sample standard deviation (SD) obtained from analyzing data generated according to models M1–M5, for all considered pool sizes when $J = 100$ in the presence of measurement error ($\tau = 0.05$). Also included are the average estimated standard errors (SE) and the empirical coverage probabilities (Cov) associated with 95% Wald confidence intervals. Two model fitting procedures were implemented, the proposed methodology (MCMLE) and the analytical approach described in Section 4 (MLE), with the latter technique only being applicable for model M1.

Model		**Measure**	***c* = 1**	***c* = 2**	***c* = 3**	***c* = 4**	***c* = 5**
M1(MCMLE)	β_0	Bias(SD)	0.00(0.07)	0.00(0.05)	0.00(0.04)	0.00(0.04)	0.00(0.03)
		Cov(SE)	0.93(0.07)	0.96(0.05)	0.94(0.04)	0.91(0.04)	0.94(0.03)
	β_1	Bias(SD)	0.00(0.05)	0.00(0.04)	0.00(0.03)	0.00(0.03)	0.00(0.02)
		Cov(SE)	0.95(0.05)	0.94(0.04)	0.97(0.03)	0.93(0.03)	0.94(0.02)
	β_2	Bias(SD)	0.00(0.10)	0.00(0.07)	0.00(0.06)	0.00(0.05)	0.00(0.05)
		Cov(SE)	0.94(0.10)	0.96(0.07)	0.94(0.06)	0.93(0.05)	0.92(0.05)
M1(MLE)	β_0	Bias(SD)	0.00(0.07)	0.00(0.05)	0.00(0.04)	0.00(0.04)	0.00(0.03)
		Cov(SE)	0.94(0.07)	0.96(0.05)	0.94(0.04)	0.92(0.04)	0.95(0.03)
	β_1	Bias(SD)	0.00(0.05)	0.00(0.04)	0.00(0.03)	0.00(0.03)	0.00(0.02)
		Cov(SE)	0.95(0.05)	0.95(0.04)	0.97(0.03)	0.93(0.03)	0.95(0.02)
	β_2	Bias(SD)	0.00(0.10)	0.00(0.07)	0.00(0.06)	0.00(0.05)	0.00(0.05)
		Cov(SE)	0.94(0.10)	0.96(0.07)	0.95(0.06)	0.93(0.05)	0.93(0.05)
M2(MCMLE)	β_0	Bias(SD)	0.02(0.08)	0.00(0.06)	0.01(0.05)	0.01(0.05)	0.01(0.04)
		Cov(SE)	0.94(0.08)	0.94(0.06)	0.94(0.05)	0.93(0.05)	0.94(0.04)
	β_1	Bias(SD)	−0.01(0.05)	0.00(0.04)	0.00(0.03)	0.00(0.03)	0.00(0.03)
		Cov(SE)	0.96(0.06)	0.93(0.04)	0.96(0.03)	0.94(0.03)	0.94(0.03)
	β_2	Bias(SD)	−0.01(0.10)	0.00(0.08)	0.00(0.06)	−0.01(0.06)	0.00(0.05)
		Cov(SE)	0.95(0.10)	0.94(0.08)	0.94(0.07)	0.94(0.06)	0.94(0.05)
M3(MCMLE)	β_0	Bias(SD)	0.00(0.17)	−0.01(0.13)	0.00(0.11)	0.00(0.09)	0.00(0.09)
		Cov(SE)	0.93(0.17)	0.94(0.13)	0.94(0.11)	0.94(0.09)	0.94(0.08)
	β_1	Bias(SD)	0.00(0.12)	0.00(0.09)	−0.01(0.08)	0.00(0.07)	0.00(0.06)
		Cov(SE)	0.95(0.12)	0.94(0.09)	0.95(0.08)	0.93(0.07)	0.94(0.06)
	β_2	Bias(SD)	−0.01(0.24)	0.01(0.19)	0.00(0.15)	0.00(0.13)	0.00(0.12)

Model		**Measure**	***c* = 1**	***c* = 2**	***c* = 3**	***c* = 4**	***c* = 5**
		Cov(SE)	0.95(0.24)	0.94(0.18)	0.95(0.15)	0.95(0.13)	0.95(0.12)
M4(MCMLE)	β_0	Bias(SD)	0.00(0.08)	0.00(0.06)	0.00(0.05)	0.00(0.05)	0.00(0.04)
		Cov(SE)	0.95(0.08)	0.97(0.06)	0.95(0.05)	0.93(0.04)	0.96(0.04)
	β_1	Bias(SD)	−0.01(0.06)	−0.01(0.04)	−0.01(0.04)	0.00(0.03)	0.00(0.03)
		Cov(SE)	0.95(0.05)	0.95(0.04)	0.93(0.03)	0.92(0.03)	0.95(0.03)
	β_2	Bias(SD)	−0.01(0.11)	0.00(0.08)	0.00(0.06)	−0.01(0.06)	0.00(0.05)
		Cov(SE)	0.92(0.10)	0.95(0.08)	0.94(0.06)	0.94(0.06)	0.94(0.05)
M5(MCMLE)	β_0	Bias(SD)	0.00(0.04)	0.00(0.03)	0.00(0.02)	0.00(0.02)	0.00(0.02)
		Cov(SE)	0.96(0.04)	0.94(0.02)	0.94(0.02)	0.94(0.02)	0.93(0.02)
	β_1	Bias(SD)	0.00(0.03)	0.00(0.02)	0.00(0.01)	0.00(0.01)	0.00(0.01)
		Cov(SE)	0.96(0.03)	0.96(0.02)	0.96(0.02)	0.94(0.01)	0.95(0.01)
	β_2	Bias(SD)	0.00(0.06)	0.00(0.04)	0.00(0.04)	0.00(0.03)	0.00(0.03)
		Cov(SE)	0.95(0.06)	0.95(0.04)	0.93(0.03)	0.95(0.03)	0.94(0.03)

Table 3

Data application: The top five models according to their AIC values for all 12 considered scenarios, where a ✓ denotes that a particular predictor was included in the model. The predictor variables x_1, x_2, and x_3 represent the participant’s age, race, and miscarriage status, respectively. Note, the presence (absence) of “ *” indicates that the MCMLE accounted for (ignored) the effect of measurement error.

Model	Data set	x_1	x_2	x_3	x_1^2	x_1x_2	x_1x_3	x_2x_3
Log-normal	ID	✓	✓	✓				✓
		✓		✓				✓
		✓	✓	✓				
		✓	✓	✓		✓		✓
		✓	✓	✓	✓			✓
	ID*	✓		✓				✓
		✓	✓	✓				✓
		✓		✓			✓	✓
		✓		✓		✓		✓
		✓	✓					
	PD			✓				✓
		✓		✓				✓
		✓		✓			✓	✓
		✓		✓		✓		✓
				✓	✓			✓
	PD*	✓		✓				✓
		✓	✓	✓				✓
		✓		✓			✓	✓
		✓	✓	✓			✓	✓
		✓	✓	✓		✓		✓
Gamma(log link)	ID	✓		✓				✓
		✓	✓	✓				✓
		✓		✓			✓	✓

Model	Data set	x_1	x_2	x_3	x_1^2	x_1x_2	x_1x_3	x_2x_3
		✓		✓	✓			✓
		✓		✓		✓		✓
	ID*	✓		✓				✓
		✓	✓	✓				✓
		✓		✓			✓	✓
		✓		✓		✓		✓
		✓		✓	✓			✓
	PD	✓	✓	✓				✓
		✓	✓	✓			✓	✓
		✓	✓	✓		✓		✓
		✓	✓	✓	✓			✓
		✓	✓	✓		✓	✓	✓
	PD*	✓	✓	✓				✓
		✓	✓	✓		✓		✓
		✓	✓	✓			✓	✓
			✓	✓		✓		✓
Gamma(inverse link)	ID	✓	✓	✓	✓	✓		✓
		✓		✓				✓
		✓	✓	✓				✓
		✓		✓			✓	✓
		✓		✓		✓		✓
	ID*	✓		✓	✓			✓
		✓		✓				✓
		✓		✓			✓	✓
		✓		✓		✓		✓
		✓	✓	✓				✓
		✓		✓	✓			✓
		✓	✓	✓				✓

Model	Data set	x_1	x_2	x_3	x_1^2	x_1x_2	x_1x_3	x_2x_3
	PD	✓	✓	✓			✓	✓
		✓	✓	✓		✓		✓
		✓	✓	✓	✓			✓
			✓	✓		✓	✓	✓
		✓	✓	✓				✓
	PD*	✓	✓	✓			✓	✓
		✓	✓	✓		✓		✓
		✓	✓	✓	✓			✓
			✓	✓		✓	✓	✓

Table 4

Data application: Presented results include the regression parameter estimates (Est.), the estimated standard errors (SE), and the corresponding p-values (P-value), under all 12 considered scenarios. Also included are the analogous results obtained from the artificial constructed homogeneous pooling data (HPD). Note, the presence (absence) of “ *” indicates that the MCMLE accounted for (ignored) the effect of measurement error.

Model	Data set		Intercept	Age	Miscarriage	Race*Miscarriage
Lognormal	ID	Est.(SE)	−2.130(0.055)	0.111(0.039)	0.271(0.087)	−0.508(0.122)
		P-value	0.000	0.005	0.002	0.000
	ID*	Est(SE)	−2.091(0.053)	0.117(0.037)	0.220(0.082)	−0.409(0.119)
		P-value	0.000	0.002	0.007	0.001
	PD	Est.(SE)	−2.438(0.070)	0.105(0.066)	0.314(0.116)	−0.822(0.225)
		P-value	0.000	0.112	0.007	0.000
	PD*	Est.(SE)	−2.365(0.077)	0.155(0.065)	0.262(0.115)	−0.681(0.246)
		P-value	0.000	0.018	0.023	0.006
	HPD*	Est.(SE)	−2.116(0.059)	0.110(0.041)	0.220(0.089)	−0.360(0.128)
		P-value	0.000	0.007	0.014	0.005
Gamma (log link)	ID	Est.(SE)	−1.592(0.075)	0.099(0.054)	0.227(0.119)	−0.406(0.167)
		P-value	0.000	0.066	0.056	0.015
	ID*	Est.(SE)	−1.615(0.053)	0.100(0.038)	0.232(0.084)	−0.403(0.119)
		P-value	0.000	0.008	0.006	0.000
	PD	Est.(SE)	−1.680(0.061)	0.171(0.058)	0.305(0.099)	−0.976(0.207)
		P-value	0.000	0.003	0.002	0.000
	PD*	Est.(SE)	−1.708(0.067)	0.176(0.063)	0.311(0.109)	−1.033(0.249)
		P-value	0.000	0.005	0.004	0.000
	HPD*	Est.(SE)	−1.601(0.058)	0.097(0.042)	0.230(0.091)	−0.407(0.130)
		P-value	0.000	0.020	0.012	0.002
Gamma (inverse link)	ID	Est.(SE)	4.922(0.365)	−0.424(0.227)	−0.959(0.511)	1.947(0.892)
		P-value	0.000	0.062	0.061	0.029
	ID*	Est.(SE)	5.033(0.265)	−0.437(0.163)	−0.989(0.369)	1.972(0.647)
		P-value	0.000	0.008	0.007	0.002
	PD	Est.(SE)	5.365(0.322)	−0.667(0.228)	−1.301(0.447)	6.429(1.915)
		P-value	0.000	0.003	0.004	0.001

Model	Data set		Intercept	Age	Miscarriage	Race*Miscarriage
	PD*	Est.(SE)	5.518(0.368)	−0.703(0.253)	−1.350(0.503)	7.002(2.507)
		P-value	0.000	0.006	0.007	0.005
	HPD*	Est.(SE)	4.964(0.286)	−0.418(0.177)	−0.977(0.398)	1.973(0.705)
		P-value	0.000	0.018	0.014	0.005