Published in final edited form as:

Stat Med. 2012 December 20; 31(29): 3896–3906. doi:10.1002/sim.5473.

A conditional likelihood approach for regression analysis using biomarkers measured with batch-specific error

Ming Wang^a, W. Dana Flanders^{a,b,c}, Roberd M. Bostick^{b,c}, and Qi Long^{a,c,*†}

^aDepartment of Biostatistics and Bioinformatics, Emory University Rollins School of Public Health, Atlanta, GA, U.S.A

^bDepartment of Epidemiology, Emory University Rollins School of Public Health, Atlanta, GA, U.S.A

^cEmory Winship Cancer Institute, Atlanta, GA, U.S.A

Abstract

Measurement error is common in epidemiological and biomedical studies. When biomarkers are measured in batches or groups, measurement error is potentially correlated within each batch or group. In regression analysis, most existing methods are not applicable in the presence of batch-specific measurement error in predictors. We propose a robust conditional likelihood approach to account for batch-specific error in predictors when batch effect is additive and the predominant source of error, which requires no assumptions on the distribution of measurement error. While a regression model with batch as a categorical covariable yields the same parameter estimates as the proposed conditional likelihood approach for linear regression, this result does not hold in general for all generalized linear models, in particular, logistic regression. Our simulation studies show that the conditional likelihood approach achieves better finite sample performance than the regression calibration approach or a naive approach without adjustment for measurement error. In the case of logistic regression, our proposed approach is shown to also outperform the regression approach with batch as a categorical covariate. In addition, we also examine a “hybrid” approach combining the conditional likelihood method and the regression calibration method, which is shown in simulations to achieve good performance in the presence of both batch-specific and measurement-specific error. We illustrate our method using data from a colorectal adenoma study.

Keywords

Batch-specific error; Biomarker; Conditional likelihood; Exponential family; Generalized linear models; Robust method

1. Introduction

It is common to observe measurement error in epidemiological and biomedical studies. In regression analysis, it is important to adjust for measurement error in predictors because failure to do so can lead to biased estimates of association or effect [1]. As a result, research on measurement error has drawn considerable interest, for which many statistical methods have been developed, e.g., regression calibration [2, 3], semiparametric methods such as sufficient and conditional scores [4, 5], and simulation extrapolation methods [6]. Carroll

et.al [7] provides a nice review of current literature on the issue of measurement error in predictors. We note that most existing methods assume a particular error structure, say additive, and a specific distribution for measurement error. While some methods avoid these assumptions, they require external/internal validation data [8, 9] or can be difficult to implement for a continuous exposure.

An important and frequent situation arises if the independent variables, either the exposure of interest or biomarkers of interest, are measured in batches or groups, which may occur in the laboratory when groups of samples are assayed together [10, 11]. This situation can result in potentially strong correlation between measurement error within each batch; we refer to this type of error as batch-specific error. Another situation in which measurement error may be correlated within groups arises if the variable has a strong temporal component, such as might occur seasonally for vitamin D or with circadian rhythm for hormone levels, assuming the long run average as the underlying exposure of interest [12].

It is a common practice to control for batch effect in the analysis of studies in which one or more independent variables are measured in groups or batches and measurement error is related to batches. As noted Blanck in a recent review [13], “If all samples cannot be assessed in a single run, then cases and an appropriate number of controls should be analyzed together in the same batch to ensure the validity of the paired comparison.” Despite this common perception and practice, the assumptions under which the paired comparisons will remove or control the batch effect are unclear. In addition, in regression analysis, most existing methods are not applicable in the presence of batch-specific error except for Long et al. [14]. Furthermore, the sensitivity of the adjustment to departures from the assumptions is also uncertain.

In this article, we present a conditional likelihood method to adjust for batch-specific error, when batch effect is additive and the predominant source of error. The method is applicable across a broad range of settings concerning the outcome-exposure relationship, specifically including the class of generalized linear models (GLMs), and is robust in that it does not require assumptions on the distribution of measurement error. The rest of the article is organized as follows. In Section 2, we describe the proposed conditional likelihood method. In Section 3, we provide simulation results to assess the finite sample performance of the proposed method under various settings. In Section 4, we illustrate the method using data from a colorectal adenoma study. We provide some concluding remarks in Section 5.

2. Methodology

2.1. The Models

Suppose that one is to estimate the association of a dependent variable y with a vector of predictors x , where x are biomarkers that are subject to batch-specific measurement error. Specifically, x is measured in G batches/groups with n_g subjects/observations in the g^{th} batch. Note that it is straightforward to extend our method to the case where only a subset of predictors are measured with batch-specific error. Here, for subject i in group g , we denote the true unobserved biomarker values by $x_{i,g}$ and the observed biomarker value by $w_{i,g}$. Assuming measurement error is additive and batch-specific, we have

$$w_{i,g} = x_{i,g} + \varepsilon_g, \quad i=1, 2, \dots, n_g; g=1, 2, \dots, G, \quad (1)$$

where ε_g is batch-specific error and is assumed to be independent of each other. Note that we make no additional assumptions about the distribution of the error terms (ε_g). We also assume the two-parameter exponential family for the dependent variable y of the form [15]:

$$f(y_{i,g}|x_{i,g}, \theta) = \exp[((\alpha + \beta^t x_{i,g})y_{i,g} - b(\alpha + \beta^t x_{i,g})) / \alpha_\phi + c_\phi(y_{i,g})] \quad (2)$$

where $y_{i,g}$ is the value of the dependent variable y for subject i in group g ; $x_{i,g} = (x_{i,g,1}, x_{i,g,2}, \dots, x_{i,g,p})^t$ is the value of the vector of predictors x for subject i in group g ; $\theta = (\alpha, \beta^t)^t$ is a $(p+1) \times 1$ vector of regression parameters to be estimated; ϕ is a dispersion parameter to encode the variance; and $a_\phi(\cdot) = a(\cdot, \phi)$, $b(\cdot)$, and $c_\phi(\cdot) = c(\cdot, \phi)$ are known smooth functions. With this formulation, the mean of y for subject i in group g , $\mu_{i,g}$, is given by

$\mu_{i,g} = \psi^{-1}(\alpha + \beta^t x_{i,g}) = \frac{db(\eta_{i,g})}{d\eta_{i,g}}$, where ψ is a link function, and $\eta_{i,g} = \alpha + \beta^t x_{i,g}$. In addition, the variance of $y_{i,g}$ is given by $V(y_{i,g}) = b''(\eta_{i,g}) a_\phi = b''(\alpha + \beta^t x_{i,g}) a_\phi$. To avoid identifiability issues, we assume the batch size $n_g > 1$ throughout this paper.

2.2. Estimation and Inference

We now present the conditional likelihood approach for estimating regression parameters $\theta = (\alpha, \beta^t)^t$ in the presence of additive batch-specific measurement error. This method conditions on group-specific totals, thus unknown measurement error is eliminated in the conditional likelihood. To fix the idea, we temporarily assume that we could observe the group-specific error (ε_g), which allows us to write the likelihood based on the true (unobserved) value of the covariates $x_{i,g} = w_{i,g} - \varepsilon_g$. The full likelihood, if the batch-specific error were known, is:

$$\begin{aligned} L &= \prod_{g=1}^G \left\{ \prod_{i=1}^{n_g} \exp \left[((\alpha + \beta^t (w_{i,g} - \varepsilon_g))y_{i,g} - b(\alpha + \beta^t (w_{i,g} - \varepsilon_g))) / \alpha_\phi + c_\phi(y_{i,g}) \right] \right\} \\ &= \prod_{g=1}^G \left\{ \exp \left[((\alpha + \beta^t (w_{1,g} - \varepsilon_g))(y_{1,g} - \sum_{j=2}^{n_g} y_{j,g}) - b(\alpha + \beta^t (w_{1,g} - \varepsilon_g))) / \alpha_\phi + c_\phi(y_{1,g} - \sum_{j=2}^{n_g} y_{j,g}) \right] \right. \\ &\quad \times \exp \left[\sum_{j=2}^{n_g} ((\alpha + \beta^t (w_{j,g} - \varepsilon_g))y_{j,g} - b(\alpha + \beta^t (w_{j,g} - \varepsilon_g))) / \alpha_\phi + c_\phi(y_{j,g}) \right] \left. \right\} \end{aligned} \quad (3)$$

Here and throughout, we use the “dot” notation to indicate summation over the missing subscript. As shown in Appendix A, unknown measurement error can be eliminated in the likelihood conditional on the group-specific totals, thus the likelihood conditional on the group-specific totals using mis-measured covariates ($w_{i,g}$) coincides with that based on the true covariates ($x_{i,g}$). The resulting conditional likelihood is given in Equation (13) in Appendix A and the conditional likelihood estimator of β can be obtained as the maximizer of the conditional likelihood. This result holds for all members of GLMs.

In the case of linear regression where $y_{i,g}$ is assumed to follow a Gaussian distribution, the conditional likelihood estimator of β , denoted by $\hat{\beta}$, has a closed-form solution

$$\hat{\beta} = \left[\sum_{g=1}^G \sum_{i=1}^{n_g} (w_{i,g} - \bar{w}_g)^t (w_{i,g} - \bar{w}_g) \right]^{-1} \sum_{g=1}^G \sum_{i=1}^{n_g} (w_{i,g} - \bar{w}_g) y_{i,g}, \quad (4)$$

where $\bar{w}_g = \frac{1}{n_g} \sum_{i=1}^{n_g} w_{i,g}$, $g = 1, 2, \dots, G$. The variance of $\hat{\beta}$ can be estimated by

$$\widehat{\text{var}}(\hat{\beta}) = \hat{\sigma}^2 / \sum_{g=1}^G \sum_{i=1}^{n_g} (w_{i,g} - \bar{w}_g)(w_{i,g} - \bar{w}_g)^t, \text{ where } \hat{\sigma}^2 \text{ can be obtained as}$$

$\left(\sum_{g=1}^G 2\widehat{t}_1 - \widehat{t}_2 - t_3 \right) / \left(\sum_{g=1}^G \frac{1-n_g^2}{n_g} \right)$ with
 $\widehat{t}_1 = \widehat{\beta}'(w_{1,g} - \bar{w}_g)y_{\cdot,g} + \sum_{j=2}^{n_g} (\widehat{\beta}'(w_{j,g} - w_{1,g}))y_{i,g}$, $\widehat{t}_2 = \sum_{i=2}^{n_g} (\widehat{\beta}'(w_{i,g} - w_{1,g}))^2$, and $t_3 = \sum_{i=1}^{n_g} (y_{i,g} - \bar{y}_g)^2$. Of note,
 the estimate based on the proposed conditional likelihood approach (4) is exactly the same
 as that from the linear regression with batch included as a categorical covariate; however,
 this result does not hold in general for all GLMs, in particular, logistic regression.

In the case of logistic regression where $y_{i,g}$ is assumed to follow a Bernoulli distribution, the
 probability mass function of the total $y_{\cdot,g}$ is

$$p(y_{\cdot,g}) = \exp \left[- \sum_{i=1}^{n_g} b(\beta'(w_{1,g} - \varepsilon_g)) \right] \sum_{\substack{y_{2,g}^*, \dots, y_{n_g,g}^* \\ \sum_{i=2}^{n_g} y_{i,g}^* = y_{\cdot,g} - y_{1,g}}} \exp \left[\beta'(w_{1,g} - \varepsilon_g)y_{\cdot,g} \right] \exp \left[\sum_{i=2}^{n_g} \beta'(w_{i,g} - w_{1,g})y_{i,g}^* \right] \quad (5)$$

where $b(t) = \log(1 + \exp(t))$. Then, the conditional likelihood can be written as

$$L_{c,l} = \prod_{g=1}^G \frac{\exp \left[\sum_{i=2}^{n_g} \beta'(w_{i,g} - w_{1,g})y_{i,g} \right]}{\sum_{\substack{y_{2,g}^*, \dots, y_{n_g,g}^* \\ \sum_{i=2}^{n_g} y_{i,g}^* = y_{\cdot,g} - y_{1,g}}} \exp \left[\sum_{i=2}^{n_g} \beta'(w_{i,g} - w_{1,g})y_{i,g}^* \right]} \quad (6)$$

where the batch error ε_g is canceled out. It follows that the above conditional likelihood is
 the same as that using the true covariates $(x_{i,g})$ and the resulting score equations are

$$s_{c,l}(\beta) = \sum_{g=1}^G \sum_{i=2}^{n_g} (w_{i,g} - w_{1,g})y_{i,g} - \sum_{g=1}^G \frac{\sum_{\substack{y_{2,g}^*, \dots, y_{n_g,g}^* \\ \sum_{i=2}^{n_g} y_{i,g}^* = y_{\cdot,g} - y_{1,g}}} \exp \left[\sum_{i=2}^{n_g} \beta'(w_{i,g} - w_{1,g})y_{i,g}^* \right] \times \sum_{i=2}^{n_g} (w_{i,g} - w_{1,g})y_{i,g}^*}{\sum_{\substack{y_{2,g}^*, \dots, y_{n_g,g}^* \\ \sum_{i=2}^{n_g} y_{i,g}^* = y_{\cdot,g} - y_{1,g}}} \exp \left[\sum_{i=2}^{n_g} \beta'(w_{i,g} - w_{1,g})y_{i,g}^* \right]} \quad (7)$$

The parameter estimates $\hat{\beta}$ can be obtained as zeros of the score equations (7). In Appendix
 B, we provide the score equations (17) for logistic regression with batch included as a
 categorical covariate, which differ from the score equations (7) and hence lead to different
 regression coefficient estimates. This result is further confirmed by our simulation studies in
 Section 3.

3. Simulation Studies

3.1. Simulation Settings

We conduct simulation studies to investigate the finite sample performance of the proposed conditional likelihood approach and we consider the cases of both linear regression and logistic regression.

In the case of linear regression, given the true biomarker value $x_{i,g}$, the continuous response variable for the i^{th} observation ($i = 1, 2, \dots, n_g$) in the g^{th} batch ($g = 1, 2, \dots, 30$), denoted by $y_{i,g}$, is generated as

$$y_{i,g} = \beta_0 + \beta_1 x_{i,g} + e_{i,g},$$

where the true regression coefficients are $\beta_0 = \beta_1 = 1$, the residual term is $e_{i,g} \sim N(0, 1)$ and the batch size n_g is generated from a Poisson distribution with a mean of $\lambda = 5$ or 10. Two types of measurement error models are used to generate the observed biomarker values ($w_{i,g}$) that are subject to measurement error. The first one is $w_{i,g} = x_{i,g} + \varepsilon_g$ where ε_g is batch-specific error and $\varepsilon_g \perp x_{i,g}$; the second one is $w_{i,g} = x_{i,g} + \varepsilon_g + \delta_{i,g}$ where $\delta_{i,g}$ is measurement-specific error, $\varepsilon_g \perp \{x_{i,g}, \delta_{i,g}\}$ and $\delta_{i,g} \perp x_{i,g}$. In all cases, $\text{Cov}(\varepsilon_g, \varepsilon_{g'}) = 0$ for $g \neq g'$. Of note, the first model represents that the observations within each batch have the same measurement error, while the second one indicates that measurement error within each batch is correlated but not exactly the same as a result of measurement-specific error. For each measurement error model, we consider two set-ups: 1) x , ε and δ follow up normal distributions, i.e., $x_{i,g} \sim N(0, \sigma_x^2)$, $\varepsilon_g \sim N(0, \sigma_\varepsilon^2)$ and $\delta_{i,g} \sim N(0, \sigma_\delta^2)$ where $\sigma_x^2 = 1$, $\sigma_\varepsilon^2 = 1$ and $\sigma_\delta^2 = 0.01, 0.25$ or 1; 2) x , ε and δ are log-normal distributed, i.e., $\log(x_{i,g}) \sim N(1, \sigma_x^2)$, $\log(\varepsilon_g) \sim N(1, \sigma_\varepsilon^2)$ and $\log(\delta_{i,g}) \sim N(1, \sigma_\delta^2)$ where $\sigma_x^2 = 1$, $\sigma_\varepsilon^2 = 1$ and $\sigma_\delta^2 = 0.01, 0.25$ or 1.

In the case of logistic regression, let $y_{i,g}$ denote the binary outcome for the i^{th} subject ($i = 1, 2, \dots, n_g$) in the g^{th} batch ($g = 1, 2, \dots, 50$), where $y_{i,g} = 1$ for cases and $y_{i,g} = 0$ for controls. Let $p_{i,g} = E(y_{i,g}/x_{i,g})$ follow a logit model

$$\text{logit}(p_{i,g}) = \beta_0 + \beta_1 x_{i,g}.$$

The regression coefficients β_0 and β_1 , the covariate $x_{i,g}$ and the batch size n_g are simulated as described for the case of linear regression. Also, the measurement error model considered here is $w_{i,g} = x_{i,g} + \varepsilon_g$, where x and ε follow normal distributions, i.e., $x_{i,g} \sim N(0, 1)$ and $\varepsilon_g \sim N(0, 1)$, or log-normal distributions, i.e., $\log(x_{i,g}) \sim N(1, 1)$ and $\log(\varepsilon_g) \sim N(1, 1)$.

In both cases of linear regression and logistic regression, we compare the proposed estimator (denoted by $\hat{\beta}_{1,CL}$) with several other estimators, namely, the “gold standard” estimator (denoted by $\hat{\beta}_{1,G}$) which uses the underlying true predictor ($x_{i,g}$), a naive estimator (denoted by $\hat{\beta}_{1,N}$) which is obtained using $w_{i,g}$ with no additional adjustment, the estimator adjusted for the batch variable (denoted by $\hat{\beta}_{1,BA}$). Specifically for linear regression, we also consider the regression calibration estimator (denoted by $\hat{\beta}_{1,RC}$) and a “hybrid” estimator that further adjusts the conditional likelihood estimator through the regression calibration approach and is denoted by $\hat{\beta}_{1,CR}$. Intuitively, the “hybrid” method may improve the performance of the conditional likelihood estimator in the presence of both batch-specific error (ε_g) and measurement-specific error ($\delta_{i,g}$) in predictors. Of note, the regression calibration coefficient

λ cannot be estimated from the data in our simulation since there are no replicate/validation data; as a result, for our cases, λ is assumed to be known exactly and can be calculated by the true variances, σ_x^2 , σ_ε^2 and σ_δ^2 . Thus, the standard error estimate for $\hat{\beta}_{1,RC}$ is computed using $SE(\hat{\beta}_{1,N})/\lambda$, where $SE(\hat{\beta}_{1,N})$ is the standard error estimate of $\hat{\beta}_1$ obtained from the naive approach and $\lambda = \sigma_x^2/(\sigma_x^2 + \sigma_\varepsilon^2 + \sigma_\delta^2)$. To compute the standard error estimate for $\hat{\beta}_{1,CL}$ in linear models, we use Equation (8); along similar lines, the standard error estimate for $\hat{\beta}_{1,CR}$ is computed using $SE(\hat{\beta}_{1,CL})/\lambda$, where $SE(\hat{\beta}_{1,CL})$ is the standard error estimate of $\hat{\beta}_1$ obtained from our conditional likelihood method and $\lambda = \sigma_x^2/(\sigma_x^2 + \sigma_\varepsilon^2)$. For each simulation set-up, 500 Monte Carlo data sets are generated with the same number of batches as the original one. The simulation results are summarized based on several measures: relative bias (*RB*), mean of standard errors (*SE*), standard deviation of the parameter estimates (*SD*), mean squared errors (*MSE*), and coverage rate of 95% confidence intervals (*CR*).

3.2. Simulation Results

For linear regression, $\hat{\beta}_{1,BA}$ is the same as $\hat{\beta}_{1,CL}$ and hence is omitted in Tables 1, 3 and 4. In all situations considered (Tables 1–4), the naive approach that does not adjust for measurement error has the worst performance as evidenced by substantial bias and the worst coverage rates that are well below the nominal level and, in particular, its point estimates bias towards the null. The “gold standard” estimator $\hat{\beta}_{1,G}$ using x without batch-specific error as the predictor clearly performs the best with negligible bias and coverage rates close to the nominal level.

Table 1 presents the simulation results for the cases where y is Gaussian and x is measured with only the batch-specific error ε_g . It shows that $\hat{\beta}_{1,CL}$ performs reasonably well when the error distribution is Gaussian or log-normal. Its performance improves as the batch size increases; in particular, when the batch size is generated from a Poisson distribution with $\lambda = 10$, its performance is comparable to that of $\hat{\beta}_{1,G}$. On the other hand, the regression calibration estimator $\hat{\beta}_{1,RC}$ exhibits moderate to considerable bias, its coverage rate is well below the nominal level, and its *MSE* is substantially larger than that of $\hat{\beta}_{1,CL}$; its performance deteriorates further when the error distribution is log-normal.

Table 2 provides the results for logistic regression. In all settings, the performance of $\hat{\beta}_{1,CL}$ is comparable to that of $\hat{\beta}_{1,G}$, whereas $\hat{\beta}_{1,N}$ performs the worst. In addition, $\hat{\beta}_{1,BA}$, different from $\hat{\beta}_{1,CL}$, exhibits substantial bias when the batch size is small and considerable bias when the batch size is moderate, and its bias is in the opposite direction of that of the naive method.

Tables 3–4 summarize additional simulation results for the case of linear regression in the presence of both batch-specific and measurement-specific error. In Table 3, x , ε and δ are normally distributed, whereas in Table 4, x , ε and δ follow log-normal distributions. In each table, three scenarios are evaluated, one with $\sigma_\varepsilon^2/(\sigma_\varepsilon^2 + \sigma_\delta^2) = 0.99$, where batch-specific error is the predominant source of error, and the other two with $\sigma_\varepsilon^2/(\sigma_\varepsilon^2 + \sigma_\delta^2) = 0.8$ and $\sigma_\varepsilon^2/(\sigma_\varepsilon^2 + \sigma_\delta^2) = 0.5$, respectively, where the amount of measurement-specific error ranges from moderate to considerable relative to batch-specific error. In all settings, again, the naive estimator exhibits substantial bias and the regression calibration estimator exhibits moderate to considerable bias and large *MSE*. $\hat{\beta}_{1,CL}$ performs reasonably well when batch-specific error is the predominant source of error and when the error distribution is either normal or log-normal; however, as the amount of measurement-specific error increases relative to batch-specific error as represented in the last two scenarios, the performance of $\hat{\beta}_{1,CL}$ deteriorates and its bias becomes substantial. At the same time, the “hybrid” estimator ($\hat{\beta}_{1,CR}$) exhibits negligible bias in all three scenarios (Tables 3 and 4); its overall

performance remains reasonably well when the error distribution is Gaussian, but the performance of its SE and the resulting confidence interval deteriorates somewhat when the error distribution is log-normal (Table 4). In Tables 3 and 4, the performance of the regression calibration estimator, again, is not satisfactory in all scenarios.

In summary, when batch-specific error is the only source of error for x or dominates measurement-specific error, the proposed conditional likelihood estimator $\hat{\beta}_{1,CL}$ performs substantially better than $\hat{\beta}_{1,N}$ and $\hat{\beta}_{1,RC}$ even if the error distribution is not Gaussian. However, its performance deteriorates in the presence of measurement-specific error in addition to batch-specific error. In the presence of both batch-specific and measurement-specific error, the “hybrid” estimator $\hat{\beta}_{1,CR}$ is preferred when the regression calibration coefficient is known or could be estimated by external or replicate data.

4. Data Example

We illustrate the proposed approach using the Markers of Adenomatous Polyps II (MAP II) study, a pilot case-control study [10, 11]. This study enrolls adult subjects without a history of previous colorectal adenoma for elective outpatient colonoscopy at a large private practice gastroenterology group to investigate possible biomarkers of risk for colorectal cancer. Biopsies from normal-appearing rectal mucosa for measurements of biomarkers were immunohistochemically processed. The subjects with adenoma are considered as cases ($y = 1$) and those without adenoma are considered controls ($y = 0$).

In this study, several plausible protein biomarkers that describe molecular phenotypes of the normal-appearing colorectal epithelium are developed and evaluated, e.g., TGF- α . Tissue samples are first immunohistochemically processed to identify biomarkers and then biomarkers are measured along the length of colon crypts, microscopic structures in the colon mucosa, using quantitative image analysis. The mean labeling optical density (expression) of each biomarker for each participant is calculated by summing the biomarker expression for all analyzed crypts and dividing by the total number of analyzed crypts. Due to limitation of the staining process, only 40 tissue samples can be processed at the same time in one batch. It is well known that measurement error in such experiments are primarily the result of the immunohistochemical process, which is affected by the lab/experiment conditions; since these conditions likely remain stable within each batch, it is reasonable to assume that measurement error in these biomarker measurements is predominantly batch-specific.

In our analysis, we consider one biomarker, TGF- α , as the predictor of interest and we are interested in assessing its association with the outcome of interest, i.e., the case/control status, through logistic regression. A total of 31 cases and 31 controls from 13 batches are used in the data analysis. For TGF- α , the biomarker values range from 0.98 to 760.78, and the mean for each batch ranges from 21.67 to 580.01. More importantly, variation observed between batches is substantially larger than variation observed within each batch, indicating that batch-specific error is likely present. We analyze the data using the proposed conditional likelihood approach as well as the naive method without adjustment for measurement error and the logistic regression model with batch as a categorical covariate. Of note, as a result of lack of replicate/validation data, the regression calibration method and the “hybrid” method are not applicable in this data analysis.

Table 5 summarizes the results for the MAP II study. The association between the case/control status and the biomarker TGF- α value is statistically significant based on the results using the proposed CL method as well as the BA method. However, the BA method results in OR that is further away from the null ($OR = 1$) and larger SE compared to the CL method, which is consistent with the simulation results. On the other hand, when using the naive

method without adjustment for measurement error, the estimated association is not statistically significant. The results again suggest that batch-specific error biases the regression coefficient estimate towards the null, illustrating the importance of adjusting for batch in the analysis or, perhaps, otherwise in the design.

5. Discussion

We propose a conditional likelihood approach for assessing associations between an outcome of interest and biomarkers measured with batch-specific error and demonstrate its usefulness and validity via theoretical arguments and numerical studies. The proposed approach yields consistent estimators without the need of correctly specifying the distribution of batch-specific error. As shown in simulations, the conditional likelihood estimator in most cases performs nearly as well as the “gold standard” estimator where no measurement error occurs. It is noted that in the case of linear regression the conditional likelihood approach is equivalent to the regression approach with batch as a categorical covariate, but this result does not hold in general for all generalized linear models. In the case of logistic regression, our simulation shows that the proposed method performs considerably better than the regression approach with batch as a categorical covariate when the batch size is small or moderate. Furthermore, as suggested by one reviewer, additional simulation studies are performed to investigate a regression approach that includes batch as random effect and this approach is shown to exhibit substantial bias and underperform the proposed conditional likelihood approach.

In this article, we formulate our approach under the framework of generalized linear models, assuming that batch effect is additive and the predominant source of error; in practice, caution is needed when using this method. For example, polynomial terms could be considered in regression models; measurement error may be multiplicative, in which case log-transformation may be needed. Furthermore, in the presence of measurement error from multiple sources (e.g., batch-specific and measurement-specific error) some “hybrid” approach may be preferred. For example, in the presence of both batch-specific and measurement-specific error as described in our simulations, one could use the “hybrid” estimator $\hat{\beta}_{1,CR}$, a conditional likelihood estimator that is further adjusted by regression calibration; however, to perform this adjustment in practice, the variance of measurement-specific error needs to be estimated from either within-batch replicate data or external validation data. The validity and operating characteristics of the “hybrid” approach need to be investigated further in future research.

Acknowledgments

This work was partly supported by NIH/NCI grant CA114456 and NIH PHS Grant UL1 RR025008 from the Clinical and Translational Science Award program. We would like to thank an Associate Editor and two referees for their helpful suggestions that improved an earlier draft of this manuscript.

References

1. Carroll, RJ. Encyclopedia of Biostatistics. 2005. Measurement error in epidemiologic studies.
2. Carroll RJ, Stefanski LA. Approximate quasilielihood estimation in models iwth surrogate predictors. Journal of the American Statistical Association. 1990; 85:652–663.
3. Glesjer, LJ. Statistical Analysis of Measurment Error Models. Providence, RI: American Mathematics Society; 1990. Improvements of the naive approach to estimation in nonlinear errors-in-variables regression models.
4. Stefanski LA, Carroll RJ. Conditional scores and optimal scores in generalized linear measurement error models. Biometrika. 1987; 74:703–716.

5. Carroll RJ, Wang MP. Semiparametric estimation in logistic measurement error models. *Journal of Royal Statistical Society Series B*. 1991; 53:573–585.
6. Cook J, Stefanski LA. A simulation extrapolation method for parametric measurement error models. *Journal of the American Statistical Association*. 1994; 89:1314–1328.
7. Carroll, RJ.; Ruppert, D.; Stefanski, LA.; Crainiceanu, CM. *Measurement error in nonlinear models: a modern perspective*. Chapman and Hall, CRC; New York: 2006.
8. Robins JM, Rotnitzky A, Zhao LP. Estimation of regression coefficients when some regressors are not always observed. *Journal of American Statistical Association*. 1994; 89:846–866.
9. Flanders WD, Greenland S. Analytic methods for two-stage case-control studies and other stratified designs. *Statistics in Medicine*. 1991; 10(5):739–747. [PubMed: 2068427]
10. Daniel CR, Bostick RM, Flanders WD, Long Q, Fedirko V, Sidelnikov E, Seabrook ME. TGF- α expression as a potential biomarker of risk within the normal-appearing colorectal mucosa of patients with and without incident sporadic adenoma. *Cancer Epidemiology Biomarkers and Prevention*. 2009; 18(1):65–73.
11. Fedirko V, Bostick RM, Flanders WD, Long Q, Shaikat A, Rutherford RE, Daniel CR, Cohen V, Dash C. Effects of vitamin D and calcium supplementation on markers of apoptosis in normal colon mucosa: a randomized, double-blind, Placebo-controlled clinical trial. *Cancer Prev Res*. 2009; 2(3):213–223.
12. Wang Y, Jacobs EJ, McCullough ML, Rodriguez C, Thun MJ, Calle EE, Flanders WD. Comparing methods for accounting for seasonal variability in a biomarker when only a single sample is available: insights from simulations based on serum 25-hydroxyvitamin d. *Am J Epidemiol*. 2009; 170(1):88–94. [PubMed: 19406919]
13. Blanck HM, Bowman BA, Cooper GR, Myers GL, Miller DT. Laboratory issues: use of nutritional biomarkers. *The Journal of Nutrition*. 2003; 133(S3):888S–894S.
14. Long Q, Flanders WD, Fedirko V, Bostick RM. Robust statistical methods for analysis of biomarkers measured with batch/experiment-specific errors. *Statistics in Medicine*. 2010; 29(3): 361–370. [PubMed: 20020422]
15. McCullagh, P.; Nelder, J. *Generalized Linear Models*. Chapman and Hall, CRC; New York: 1989.

Appendix A

For a generalized linear model, we here derive an expression for the conditional likelihood given the group-specific sum $y_{\cdot g}$. For simplicity, we temporarily drop the subscript g , and consider the following change of variables: $s_1 = y_{\cdot}$, and $s_j = y_j$ for $j = 2, \dots, n$. The inverse transformation is $y_1 = s_1 - s_2 - \dots - s_n$; and $y_j = s_j$ for $j = 2, \dots, n$. The determinant of the Jacobian is 1. We take $f_j(s_j)$ as the density function of s_j with respect to the dominating measure ζ , where ζ is Lebesgue measure on $\mathbb{R}$ or counting measure on $\mathbb{N}$.

In general, the change of variable technique implies that the distribution of $s_1 = y_{\cdot}$ is given by:

$$L_{s_1} = \int \cdots \int f_1(s_1 - s_2 - \dots - s_n) f_2(s_2) \cdots f_n(s_n) d\zeta(s_2) \cdots d\zeta(s_n) \quad (8)$$

Here the integration is over the collection of all $s_2, s_3, \dots, s_n$ such that would be in the range of permitted values for y_1 . The expression (8) for the exponential family is given by:

$$\begin{aligned} L_{s_1} = & \int \cdots \int \left\{ \exp \left[((\alpha + \beta^t(w_1 - \varepsilon))(s_1 - \sum_{j=2}^n s_j) - b(\alpha + \beta^t(w_1 - \varepsilon))) / \alpha_\phi + c_\phi(s_1 - \sum_{j=2}^n s_j) \right] \right. \\ & \times \exp \left[\sum_{j=2}^n (\alpha + \beta^t(w_j - \varepsilon)s_j - b(\alpha + \beta^t(w_j - \varepsilon))) / \alpha_\phi + c_\phi(s_j) \right] \left. \right\} d\zeta(s_2) \cdots d\zeta(s_n) \end{aligned} \quad (9)$$

The terms $(\alpha + \beta^t \varepsilon)s_j$ for $j = 2, \dots, n$ are canceled from expression (9) leaving after rearrangement:

$$L_{s_1} = \int \cdots \int \left\{ \exp[(\alpha + \beta^t(w_1 - \varepsilon)s_1 - b(\alpha + \beta^t(w_1 - \varepsilon)))/\alpha_\phi + c_\phi(s_1 - \sum_{j=2}^n s_j)] \right. \\ \left. \times \exp[\sum_{j=2}^n (\alpha + \beta^t(w_j - w_1)s_j - b(\alpha + \beta^t(w_j - \varepsilon)))/\alpha_\phi + c_\phi(s_j)] \right\} d\zeta(s_2) \cdots d\zeta(s_n) \quad (10)$$

Now the desired conditional likelihood is the ratio of expression (3) in the main text after the change of variables divided by expression (10). In this ratio, the terms involving $b(\alpha + \beta^t(w_j - \varepsilon))$ appear as factors, and can be canceled from numerator and denominator. Canceling these common terms gives the contribution to the likelihood from group g :

$$L_c = \frac{\exp[(\beta^t(w_1 - \varepsilon)y)/\alpha_\phi + c_\phi(y - \sum_{j=2}^n y_j) + \sum_{j=2}^n (\beta^t(w_j - w_1)y_j)/\alpha_\phi + c_\phi(y_j)]}{\int \cdots \int \left\{ \exp[(\beta^t(w_1 - \varepsilon)s_1)/\alpha_\phi + c_\phi(s_1 - \sum_{j=2}^n s_j) + \sum_{j=2}^n (\beta^t(w_j - w_1)s_j)/\alpha_\phi + c_\phi(s_j)] \right\} d\zeta(s_2) \cdots d\zeta(s_n)} \quad (11)$$

Now the coefficient of ε in the numerator is $\beta^t \sum_{j=1}^n y_j / \alpha_\phi$, and remembering that $s_1 = y$, and that integration is not over s_1 , we see that the last term involving the unknown ε cancels from numerator and denominator leaving, after restoring the dependency on g :

$$L_{c,g} = \frac{\exp[(\beta^t w_{1,g} y_{\cdot,g})/\alpha_\phi + c_\phi(y_{\cdot,g} - \sum_{j=2}^{n_g} y_{j,g}) + \sum_{j=2}^{n_g} (\beta^t(w_{j,g} - w_{1,g})y_{j,g})/\alpha_\phi + c_\phi(y_{j,g})]}{\int \cdots \int \left\{ \exp[(\beta^t(w_{1,g} s_{1,g})/\alpha_\phi + c_\phi(s_{1,g} - \sum_{j=2}^{n_g} s_{j,g}) + \sum_{j=2}^{n_g} (\beta^t(w_{j,g} - w_{1,g})s_{j,g})/\alpha_\phi + c_\phi(s_{j,g})] \right\} d\zeta(s_{2,g}) \cdots d\zeta(s_{n_g,g})} \quad (12)$$

The conditional likelihood is then:

$$L_c = \prod_{g=1}^G L_{c,g} \quad (13)$$

To obtain the parameter estimates based on (13), the key step is to calculate the integral in the denominator of (12). A close form of this integral can be obtained in linear regression as indicated in Section 2.2, but it is more involved to compute this integral for other GLMs. Numerical evaluation of the integral can be performed for discrete outcomes and a Gauss-Hermite quadrature method can be used to approximate the integral numerically for continuous outcomes.

Appendix B

For the Binomial models with batch as a discrete covariate, we know the log-likelihood is

$$l(\beta) = \sum_{g=1}^G \sum_{i=1}^{n_g} \left[y_{i,g} \log(\pi_{i,g}) + (1-y_{i,g}) \log(1-\pi_{i,g}) \right] \quad (14)$$

where $\pi_{i,g} = \frac{\exp(\beta_g + \beta_x^t x_{i,g})}{1 + \exp(\beta_g + \beta_x^t x_{i,g})}$. Thus, for $g = 1, 2, \dots, G$, we get the score equation

$$\begin{aligned} \frac{\partial l}{\partial \beta_g} &= \sum_{i=1}^{n_g} \left\{ \frac{y_{i,g}}{1 + \exp(\beta_g + \beta_x^t x_{i,g})} - \frac{(1-y_{i,g}) \exp(\beta_g + \beta_x^t x_{i,g})}{1 + \exp(\beta_g + \beta_x^t x_{i,g})} \right\} \\ &= \sum_{i=1}^{n_g} \left\{ y_{i,g} - \frac{\exp(\beta_g + \beta_x^t x_{i,g})}{1 + \exp(\beta_g + \beta_x^t x_{i,g})} \right\} = 0 \end{aligned}$$

Thus, from (15), we can get that

$$\exp(\beta_g) = \frac{\sum_{i=1}^{n_g} y_{i,g}}{\sum_{i=1}^{n_g} \exp(\beta_x^t x_{i,g}) (1-y_{i,g})} \quad (15)$$

Similarly, for the parameter β_x , we can get the partial score

$$\begin{aligned} \frac{\partial l}{\partial \beta_x^t} &= \sum_{g=1}^G \sum_{i=1}^{n_g} \left\{ \frac{y_{i,g}}{1 + \exp(\beta_g + \beta_x^t x_{i,g})} - \frac{(1-y_{i,g}) \exp(\beta_g + \beta_x^t x_{i,g}) x_{i,g}}{1 + \exp(\beta_g + \beta_x^t x_{i,g})} \right\} \\ &= \sum_{g=1}^G \sum_{i=1}^{n_g} \left\{ \frac{y_{i,g} [1 + \exp(\beta_g + \beta_x^t x_{i,g}) x_{i,g}]}{1 + \exp(\beta_g + \beta_x^t x_{i,g})} - \frac{\exp(\beta_g + \beta_x^t x_{i,g}) x_{i,g}}{1 + \exp(\beta_g + \beta_x^t x_{i,g})} \right\} \end{aligned}$$

Thus, from (16), we can get that

$$\sum_{g=1}^G \sum_{i=1}^{n_g} y_{i,g} - \sum_{g=1}^G \sum_{i=1}^{n_g} \left\{ \exp(\beta_g + \beta_x^t x_{i,g}) (1-y_{i,g}) x_{i,g} \right\} = 0 \quad (16)$$

Plug the results from (15) into (16), we have

$$\sum_{g=1}^G \sum_{i=1}^{n_g} y_{i,g} - \sum_{g=1}^G \sum_{i=1}^{n_g} \left\{ \frac{\sum_{i=1}^{n_g} y_{i,g} \exp(\beta_x^t x_{i,g}) (1-y_{i,g}) x_{i,g}}{\sum_{i=1}^{n_g} \exp(\beta_x^t x_{i,g}) (1-y_{i,g})} \right\} = 0 \quad (17)$$

Then the estimate of the parameters β_x can be solved from Equation (17).

Table 1

Linear regression model with true $\beta_1 = 1$ and additive error, $W_{i,g} = x_{i,g} + \varepsilon_g$, $N = 30$ batches.

	Poisson parameter $\lambda = 5$					Poisson parameter $\lambda = 10$				
	RB (%)	SE	SD	MSE	CR (%)	RB (%)	SE	SD	MSE	CR (%)
$x_{i,g} \sim N(0, \sigma_x^2)$ and $\varepsilon_g \sim N(0, \sigma_\varepsilon^2)$ with $\sigma_x^2 = 1$ and $\sigma_\varepsilon^2 = 1$										
$\hat{\beta}_{1,G}$	-0.07	0.082	0.088	0.088	93.0	-0.03	0.058	0.056	0.056	95.2
$\hat{\beta}_{1,N}$	-48.8	0.072	0.098	0.498	0.2	-48.5	0.050	0.082	0.492	0.0
$\hat{\beta}_{1,CL}$	0.12	0.079	0.096	0.096	90.6	-0.02	0.060	0.060	0.060	94.0
$\hat{\beta}_{1,RC}$	2.3	0.144	0.197	0.198	84.6	2.9	0.101	0.162	0.164	75.6
$\log(x_{i,g}) \sim N(1, \sigma_x^2)$ and $\log(\varepsilon_g) \sim N(1, \sigma_\varepsilon^2)$ with $\sigma_x^2 = 1$ and $\sigma_\varepsilon^2 = 1$										
$\hat{\beta}_{1,G}$	-0.09	0.087	0.093	0.093	93.8	0.09	0.059	0.057	0.057	95.8
$\hat{\beta}_{1,N}$	-44.1	0.074	0.192	0.481	5.8	-43.0	0.052	0.167	0.461	1.2
$\hat{\beta}_{1,CL}$	0.09	0.084	0.104	0.104	89.6	-0.05	0.063	0.061	0.061	95.4
$\hat{\beta}_{1,RC}$	11.7	0.149	0.383	0.401	52.4	13.9	0.104	0.335	0.362	39.6

RB, relative bias; SE, mean of standard errors; SD, standard deviation of parameter estimates; MSE, mean squared errors; CR, coverage rate of 95% confidence intervals.

Table 2

Logistic regression model with true $\beta_1 = 1$ and additive error, $W_{i,g} = x_{i,g} + \varepsilon_g$, $N = 30$ batches.

	Poisson parameter $\lambda = 5$					Poisson parameter $\lambda = 10$				
	RB (%)	SE	SD	MSE	CR (%)	RB (%)	SE	SD	MSE	CR (%)
$x_{i,g} \sim N(0, \sigma_x^2)$ and $\varepsilon_g \sim N(0, \sigma_\varepsilon^2)$ with $\sigma_x^2 = 1$ and $\sigma_\varepsilon^2 = 1$										
$\hat{\beta}_{1,G}$	1.39	0.179	0.172	0.172	95.0	-0.007	0.126	0.131	0.131	93.6
$\hat{\beta}_{1,N}$	-52.2	0.113	0.122	0.536	2.4	-53.5	0.078	0.094	0.544	0.0
$\hat{\beta}_{1,CL}$	2.28	0.211	0.215	0.216	95.4	-0.37	0.134	0.140	0.140	94.2
$\hat{\beta}_{1,BA}$	30.3	0.244	0.292	0.420	80.6	12.4	0.145	0.162	0.204	86.0
$\log(x_{i,g}) \sim N(1, \sigma_x^2)$ and $\log(\varepsilon_g) \sim N(1, \sigma_\varepsilon^2)$ with $\sigma_x^2 = 1$ and $\sigma_\varepsilon^2 = 1$										
$\hat{\beta}_{1,G}$	-6.9	0.401	0.423	0.428	96.0	1.81	0.275	0.274	0.275	96.8
$\hat{\beta}_{1,N}$	-60.9	0.196	0.235	0.653	21.0	-64.8	0.132	0.165	0.669	3.0
$\hat{\beta}_{1,CL}$	9.1	0.447	0.516	0.524	94.6	1.54	0.287	0.291	0.291	96.8
$\hat{\beta}_{1,BA}$	34.7	0.505	0.690	0.772	91.8	11.9	0.303	0.326	0.347	95.8

RB, relative bias; SE, mean of standard errors; SD, standard deviation of parameter estimates; MSE, mean squared errors; CR, coverage rate of 95% confidence intervals.

Table 3

Linear regression model with true $\beta_1 = 1$ and additive error, $W_{i,g} = x_{i,g} + \varepsilon_g + \delta_{i,g}; x_{i,g} \sim N(0, \sigma_x^2)$ with $\sigma_x^2 = 1; \varepsilon_g \sim N(0, \sigma_\varepsilon^2)$ with $\sigma_\varepsilon^2 = 1; \delta_{i,g} \sim N(0, \sigma_\delta^2)$. $N = 30$ batches.

Poisson parameter $\lambda = 5$						Poisson parameter $\lambda = 10$				
RB (%)	SE	SD	MSE	CR (%)	RB (%)	SE	SD	MSE	CR (%)	
$\sigma_{\varepsilon}^2/(\sigma_{\varepsilon}^2+\sigma_{\delta}^2)=0.99, \sigma_{\delta}^2=0.01$										
$\hat{\beta}_{1,G}$	-0.6	0.083	0.088	0.088	94.2	-0.5	0.058	0.056	0.056	94.0
$\hat{\beta}_{1,N}$	-48.6	0.072	0.094	0.495	0.2	-48.1	0.051	0.083	0.488	0.0
$\hat{\beta}_{1,CL}$	-1.5	0.085	0.096	0.097	91.0	-1.6	0.063	0.060	0.061	94.8
$\hat{\beta}_{1,CR}$	-0.5	0.086	0.097	0.097	91.6	-0.6	0.063	0.060	0.060	95.4
$\hat{\beta}_{1,RC}$	3.4	0.145	0.189	0.192	85.2	4.3	0.102	0.167	0.173	76.8
$\sigma_{\varepsilon}^2/(\sigma_{\varepsilon}^2+\sigma_{\delta}^2)=0.8, \sigma_{\delta}^2=0.25$										
$\hat{\beta}_{1,G}$	-0.6	0.083	0.088	0.088	94.2	-0.5	0.058	0.056	0.056	94.0
$\hat{\beta}_{1,N}$	-54.4	0.069	0.084	0.550	0.0	-53.9	0.048	0.072	0.544	0.0
$\hat{\beta}_{1,CL}$	-20.6	0.083	0.094	0.227	32.2	-20.6	0.061	0.059	0.214	7.2
$\hat{\beta}_{1,CR}$	-0.8	0.104	0.117	0.117	91.0	-0.8	0.077	0.073	0.074	96.4
$\hat{\beta}_{1,RC}$	2.6	0.156	0.188	0.190	89.6	3.6	0.109	0.162	0.166	80.0
$\sigma_{\varepsilon}^2/(\sigma_{\varepsilon}^2+\sigma_{\delta}^2)=0.5, \sigma_{\delta}^2=1$										
$\hat{\beta}_{1,G}$	-0.6	0.083	0.088	0.088	94.2	-0.5	0.058	0.056	0.056	94.0
$\hat{\beta}_{1,N}$	-66.2	0.062	0.067	0.067	0.0	-65.9	0.043	0.053	0.661	0.0
$\hat{\beta}_{1,CL}$	-50.4	0.073	0.082	0.511	0.0	-50.4	0.054	0.053	0.507	0.0
$\hat{\beta}_{1,CR}$	-0.8	0.146	0.164	0.164	92.2	-0.9	0.108	0.106	0.106	92.2
$\hat{\beta}_{1,RC}$	1.3	0.185	0.200	0.200	93.6	2.3	0.130	0.159	0.160	88.4

RB, relative bias; SE, mean of standard errors; SD, standard deviation of parameter estimates; MSE, mean squared errors; CR, coverage rate of 95% confidence intervals.

Table 4

Linear regression model with true $\beta_1 = 1$ and additive error, $W_{i,g} = x_{i,g} + \varepsilon_g + \delta_{i,g}; \log(x_{i,g}) \sim N(1, \sigma_x^2)$ with $\sigma_x^2 = 1; \log(\varepsilon_g) \sim N(1, \sigma_\varepsilon^2)$ with $\sigma_\varepsilon^2 = 1; \log(\delta_{i,g}) \sim N(1, \sigma_\delta^2)$. $N = 30$ batches.

	Poisson parameter $\lambda = 5$				Poisson parameter $\lambda = 10$					
	RB (%)	SE	SD	MSE	CR (%)	RB (%)	SE	SD	MSE	CR (%)
$\sigma_\varepsilon^2/(\sigma_\varepsilon^2+\sigma_\delta^2)=0.99, \sigma_\delta^2=0.01$										
$\hat{\beta}_{1,G}$	-0.1	0.087	0.097	0.097	91.0	-0.2	0.060	0.057	0.057	96.0
$\hat{\beta}_{1,N}$	-44.9	0.074	0.182	0.484	3.6	-43.1	0.052	0.182	0.468	1.2
$\hat{\beta}_{1,CL}$	-1.2	0.090	0.103	0.104	91.6	-1.4	0.065	0.060	0.062	95.8
$\hat{\beta}_{1,CR}$	-0.2	0.091	0.104	0.104	91.6	-0.4	0.065	0.061	0.061	97.0
$\hat{\beta}_{1,RC}$	10.7	0.149	0.365	0.380	54.0	14.3	0.104	0.366	0.393	33.0
$\sigma_\varepsilon^2/(\sigma_\varepsilon^2+\sigma_\delta^2)=0.8, \sigma_\delta^2=0.25$										
$\hat{\beta}_{1,G}$	-0.1	0.087	0.097	0.097	91.0	-0.2	0.060	0.057	0.057	96.0
$\hat{\beta}_{1,N}$	-52.3	0.071	0.155	0.546	0.0	-50.6	0.050	0.153	0.529	0.0
$\hat{\beta}_{1,CL}$	-22.1	0.086	0.120	0.251	33.2	-21.4	0.063	0.077	0.228	9.2
$\hat{\beta}_{1,CR}$	-2.6	0.107	0.150	0.152	82.6	-1.8	0.078	0.097	0.098	87.2
$\hat{\beta}_{1,RC}$	7.3	0.159	0.350	0.357	59.6	11.2	0.112	0.345	0.362	40.0
$\sigma_\varepsilon^2/(\sigma_\varepsilon^2+\sigma_\delta^2)=0.5, \sigma_\delta^2=1$										
$\hat{\beta}_{1,G}$	-0.1	0.087	0.097	0.097	91.0	-0.2	0.060	0.057	0.057	96.0
$\hat{\beta}_{1,N}$	-65.2	0.063	0.128	0.664	0.0	-64.1	0.044	0.115	0.652	0.0
$\hat{\beta}_{1,CL}$	-50.4	0.074	0.150	0.526	0.2	-50.2	0.055	0.109	0.514	0.0
$\hat{\beta}_{1,CR}$	-0.8	0.149	0.299	0.299	63.2	-0.4	0.109	0.219	0.219	66.4
$\hat{\beta}_{1,RC}$	4.5	0.189	0.384	0.386	65.6	7.6	0.132	0.345	0.353	54.0

RB, relative bias; SE, mean of standard errors; SD, standard deviation of parameter estimates; MSE, mean squared errors; CR, coverage rate of 95% confidence intervals.

Table 5

The results for the MAP II study. *N*, the naive method that does not adjust for measurement error; *CL*, the proposed conditional likelihood approach; *BA*, the logistic regression including batch as a categorical covariate. *SE*, standard error; *OR*, odds ratio for 10-unit change in TGF- α ; p-value, based on likelihood ratio tests.

	$10 \times \hat{\beta}(SE)$	OR	P-value
<i>N</i>	-0.027(0.020)	0.973	0.321
<i>CL</i>	-0.077(0.040)	0.926	0.027 [*]
<i>BA</i>	-0.096(0.044)	0.908	0.025 [*]

^{*}
:significant at 0.05 level