Published in final edited form as:

Stat Med. 2010 February 10; 29(3): 361–370. doi:10.1002/sim.3796.

Robust Statistical Methods for Analysis of Biomarkers Measured with Batch/Experiment Specific Errors

Qi Long^{1,3}, W. Dana Flanders^{1,2,3}, Veronika Fedirko^{2,3}, and Roberd M. Bostick^{2,3}

¹Department of Biostatistics and Bioinformatics, Rollins School of Public Health, Emory University, Atlanta, GA, 30322, USA

²Department of Epidemiology, Rollins School of Public Health, Emory University, Atlanta, GA, 30322, USA

³Winship Cancer Institute, Emory University, Atlanta, GA, 30322, USA

SUMMARY

In many biological studies, biomarkers are measured with errors. In addition, study samples are often divided and measured in separate batches, and data collected from different experiments are used in a single analysis. Generally speaking, the structure of the measurement error is unknown and is not easy to ascertain. While the conditions under which the measurements are taken vary from one batch/experiment to another, they are often held steady within each batch/experiment. Thus, the measurement error can be considered batch/experiment specific, that is, fixed within each batch/experiment, which result into a rank preserving property within each batch/experiment. Under this condition, we study robust statistical methods for analyzing the association between an outcome variable and predictors measured with error, and evaluating the diagnostic or predictive accuracy of these biomarkers. Our methods require no assumptions on the structure and distribution of the measurement error, which are often unrealistic. Compared to existing methods that are predicated on normality and additive structure of measurement errors, our methods still yield valid inferences under departure from these assumptions. The proposed methods are easy to implement using off-shelf software. Simulation studies show that under various measurement error structures, the performance of the proposed methods is satisfactory even for a fairly small sample size, whereas existing methods under misspecified structures and a naive approach exhibited substantial bias. Our methods are illustrated using a biomarker validation case-control study for colorectal neoplasms.

Keywords

batch effect; batch/experiment specific error; measurement error; ROC analysis; surrogate variable

1. Introduction

Biomarkers are measured in many clinical trials and observational studies, and are then used to evaluate their association with risk of a disease, which is of substantial interest to researchers in many disease areas such as cancer. It is well known that biomarkers are often measured with errors in bioassays. Furthermore, when there is a need to measure biomarkers from a large number of samples in a study, the samples are often divided into several batches, which are processed and measured under different conditions at different times by

different technicians. Within each batch, the measurements are taken under similar conditions. This phenomenon is referred to as batch effects and has been previously described and discussed in the areas of cancer and nutrition research [1, 2, 3, 4]. The batch effect can also be extended to data collected from multiple experiments when conducting a meta-data analysis [5, 6].

Statistical analysis that do not account for contaminated biomarker values is not appropriate. There have been extensive discussions in current literature for measurement error problems; Carroll et al. [7] provides an excellent review of many existing methods. To name a few, Stefanski and Carroll [8] proposed a semiparametric approach based on conditional scores; Nakamura [9] and Stefanski [10] proposed another semiparametric approach based on corrected scores; Stefanski and Buzas [11] proposed an instrumental variable estimation approach; and Cook and Stefanski [12] proposed a simulation-extrapolation procedure; among others. In general, the existing statistical methods require assumptions on the structure of measurement errors, and measurement errors are assumed to be either additive or multiplicative and follow a normal distribution with mean 0 and constant variance. In practice, the validity of these assumptions are unknown to investigators and can not be empirically verified, and they are often unrealistic. For instance, in many bioassays, the amount of a protein biomarker that is present is represented using the staining density measured with a microscope, and the immunohistochemical process is the primary reason for introducing potentially substantial measurement errors and is affected by the lab/experiment conditions. While investigators attempt to control these conditions, it can be very challenging to always maintain similar conditions between multiple experiments or different batches of a same study. As a result, the error structure and distribution may change from one batch to another. To further complicate the data analysis, in most cases no alternative technology is available to validate these measurements and replicate measurements are not possible in many cases; for instance, in bioassay each tissue sample can only be stained once. Hence, it is difficult to ascertain the structure and distribution of the measurement errors. Because of these complicating issues, existing statistical methods are not applicable when data are collected from several batches in one study or from multiple experiments.

In the presence of the batch effect, alternatively it is often plausible to assume that the measurement error is roughly the same for measurements taken within the same batch/experiment, which is defined as the batch/experiment specific measurement error. As a result, within each batch the order of the measurements that are contaminated with errors is the same as that of the underlying true values, but this property may not hold across different batches/experiments. This type of measurement error structure has been previously exploited in [6], where a standardization step was carried out for each gene microarray measurement before additional data analysis was conducted. Specifically, the batch/experiment-specific mean was subtracted from each measurement, and this step was aimed to remove the noise that is reliant on the lab/experiment conditions. It was shown [6] that if combining data from two separate experiments without standardization, the gene expression data displayed no clear patterns between the diseased and non-diseased; the standardization step helped reduce the noise due to different experiments and as a result clear patterns emerged. This standardization procedure implicitly assumes that the measurement error is fixed for each batch/experiment and is additive.

In this paper, we are interested in 1) modeling the association between an outcome and an explanatory variable that is measured with batch/experiment specific errors; and 2) evaluating the diagnostic/predictive accuracy of such an explanatory variable when the outcome of interest is a disease status. We propose two robust methods which do not rely on assumptions on the structure and distribution of measurement errors. This paper is organized as follows. In Section 2, we introduce the notation used throughout the paper, and describe

the robust methods for data analysis and their theoretical properties. In Section 3, we conduct detailed simulation studies to assess the impact of several factors on the finite sample performance of the proposed methods. In Section 4, we apply the proposed methods to data from a study of biomarkers of risk for sporadic colorectal neoplasms. In Section 5, we provide some discussion and concluding remarks.

2. Methodology

2.1. Batch/Experiment Specific Measurement error

Without loss of generality, we consider a study with n observations and the study is aimed to examine the association of one outcome (Y) with one explanatory variable (X), where the explanatory variable is measured with measurement errors and in different batches. The observations within each batch are a random sample from either the entire study population, which is referred to as a cohort design, or stratified on disease status when Y is an indicator for disease status, which is referred to as a case-control design. It is straightforward to generalize to multiple explanatory variables.

For the i th observation from the b th batch, let Y_{bi} and X_{bi} ($i = 1, \dots, n_b$ and $b = 1, \dots, B$) be the outcome and the true biomarker measurement. One does not actually observe X_{bi} but instead observes

$$W_{bi} = h(X_{bi}, \eta_b) \quad (1)$$

where η_b captures the conditions of an experiment and is batch/experiment specific, and h is an unknown monotonic function in X_{bi} for given η_b . There is a subtle yet important difference between our model (1) and traditional measurement error models [7]. Under Model (1), W can be considered as a type of surrogate variable (say, the staining density) for the true variable of interest X (say, the amount of protein present in the tissue), and their relationship $h(\cdot)$ is affected by lab/experiment conditions (η_b). In order for W to be a surrogate for X , $h(\cdot)$ needs to be monotonic in X for fixed η_b and without loss of generality $h(\cdot)$ is assumed to be increasing throughout. η_b is batch/experiment specific, i.e., fixed within each batch/experiment, and it is termed batch/experiment specific error in Model (1). While it is unrealistic to assume a simple structure for $h(\cdot)$ (say, additive), it is straightforward to show that for any two observations in a same batch b , if $X_{bi} \leq X_{bj}$ then $W_{bi} \leq W_{bj}$, which is a rank preserving property within each batch. On the other hand, the measurement error defined as in traditional models [7] with a specific error structure (say, additive and $\varepsilon_{bi} = W_{bi} - X_{bi}$) can be different even within the same batch. As a special case of Model (1), a traditional additive measurement error model with batch-specific errors corresponds to $\eta_b = \varepsilon_b$ and $W_{bi} = X_{bi} + \varepsilon_b$. In this article, we study statistical methods for our general model (1) that do not rely on assumptions on the structure of h and the distribution of η_b , and that η_b has mean 0.

2.2. Assessment of Association

We propose a general approach to assess the association of Y and X under Model (1). We first define two new variables, Z_{bi} based on X and Z_{bi}^* based on W , such that $Z_{bi}^* \rightarrow_p Z_{bi}$. For example, $Z_{bi} = I(X_{bi} \leq \zeta_X)$, where ζ_X is a p th percentile such as the median of the distribution of X , and $Z_{bi}^* = I(W_{bi} \leq \hat{\zeta}_b)$, where ζ_b and $\hat{\zeta}_b$ are the true p th percentile and the sample p th percentile of W in the b th batch, respectively. Z_{bi} is not observed whereas Z_{bi}^* is. Let $\rightarrow_p$ denote convergence in probability. One can show that as $n_b \rightarrow \infty$, $Z_{bi}^* \rightarrow_p Z_{bi}$ when the rank-preserving assumption holds (Appendix A). More generally, Z_{bi} can be defined based on the

quantities of X for the entire study population and Z_{bi}^* can be defined based on the sample quantities of $\{W_{b1}, \dots, W_{bn_b}\}$, that is, within the b th batch.

Given Z_{bi}^* , we can examine the association of Z_{bi}^* and Y using an appropriate statistical method. We further assume that given X_{bi} , Y_{bi} follows a distribution from an exponential family and $\mu_{bi} = E(Y_{bi}|X_{bi})$ follows a generalized linear model

$$g(\mu_{bi}) = \alpha + \beta X_{bi} \quad (2)$$

where $g(\cdot)$ is a monotonic function. Model (2) and the distribution of X , in turn, induce a true model for Z , that is,

$$g(\mu_{bi}) = \alpha_1 + \gamma_T Z_{bi}. \quad (3)$$

We note that Z is not observed and therefore γ_T can not be estimated using observed data. Instead, using observed Z^* , we propose to fit a model

$$g(\mu_{bi}) = \alpha_1 + \gamma Z_{bi}^*. \quad (4)$$

A maximum likelihood estimate of γ in Model (4), $\hat{\gamma}$, can be computed, and it can be shown that $\hat{\gamma}$ is a consistent estimate of γ_T when $n_b \rightarrow \infty$ (Appendix A). The parameter estimates and their standard errors can be computed using off-shelf software such as SAS.

When Y is an indicator for disease status (1 for diseased subjects and 0 for non-diseased subjects) and the median is used to define Z^* , a logit model can be used for (3) and one can interpret γ as the odds of being diseased vs non-diseased for subjects with W higher than the median of W is e^γ times the odds for subjects with W lower than the median of W , that is, e^γ is an odds ratio (OR). Due to the rank-preserving property, e^γ is also the odds ratio of being diseased vs non-diseased for subjects with X higher than its median compared to subjects with X lower than its median.

Several remarks are in order. First, ζ_b needs to be estimated. Second, order statistics such as quartiles are preferred for ζ_b since they are transformation invariant under the rank-preserving property; the interpretation will change according to the definition of ζ_b . Third, ζ_b can be based on a quantity from only the diseased subjects or non-diseased subjects when Y is an indicator for diseased or nondiseased. Last, other statistical methods may be used for estimating the association between Z^* and W , say, a nonparametric estimate of an OR.

2.3. Assessment of Diagnostic/Predictive Accuracy

When Y is an indicator of disease status, investigators are often interested in the diagnostic or predictive accuracy of the true biomarker measurement X . We propose to use a stratified Mann-Whitney U statistic to estimate the ROC (Receiver Operating Characteristic) curve of using X predicting Y , which can also be used to compare the predictive accuracy of multiple biomarkers. Let X_1 and X_0 denote the biomarker variable in cases and controls, respectively, and assume that cases tend to have higher biomarker values. Let X_{1bi} ($i = 1, \dots, m_{1b}$) and X_{0bi} ($i = 1, \dots, m_{0b}$) denote the i th observed biomarker value in the b th batch for cases and controls, respectively. Similarly, we can define $W_{1bi} = g(X_{1bi}, \eta_b)$ and $W_{0bi} = g(X_{0bi}, \eta_b)$. Also let $m_1 = \sum_b m_{1b}$ denote the total number of diseased and $m_0 = \sum_b m_{0b}$ denote the total

number of nondiseased; hence $n = m_1 + m_0$. The objective is to estimate the area under the curve (AUC) of a ROC curve when using X predicts Y . It is well known that the AUC is equal to the Mann-Whitney U statistic [13], that is $\text{AUC} = \theta = \psi(X_1, X_0)$, where $\psi(X_1, X_0) = 1$, if $X_1 > X_0$, 0.5 if $X_1 = X_0$, and 0 if $X_1 < X_0$. Since X_1 and X_0 are not observed, one needs to use the observed data W_1 and W_2 to estimate θ . We propose to estimate θ as follows

$$\hat{\theta} = \sum_b w_b \hat{\theta}_b \quad (5)$$

where $w_b = (m_{1b} + m_{0b}) / (m_1 + m_0)$ is a weight function and $\hat{\theta}_b = \frac{1}{m_{1b} m_{0b}} \sum_{i=1}^{m_{1b}} \sum_{j=1}^{m_{0b}} \psi(W_{1bi}, W_{0bj})$.

If the batch/experiment specific assumption holds, it can be shown that $\hat{\theta}$ is a consistent estimator for θ (Appendix B) when $m_{1b} \rightarrow \infty$ and $m_{0b} \rightarrow \infty$. It can also be shown that the large sample variance of $\hat{\theta}$ is

$$\sum_{b=1}^B \frac{w_b^2}{m_{1b} m_{0b}} \{ \theta_b(1 - \theta_b) + (m_{1b} - 1)(R_{1b} - \theta_b^2) + (m_{0b} - 1)(R_{0b} - \theta_b^2) \} \quad (6)$$

where $R_{1b} = P(W_{1bi} > W_{0bj}, W_{1bi'} > W_{0bj})$, $R_{0b} = P(W_{1bi} > W_{0bj}, W_{1bi} > W_{0bj'})$, and $(W_{1bi}, W_{1bi'})$ and $(W_{0bj}, W_{0bj'})$ denote random pairs of observations from diseased and non-diseased populations within the b th batch, respectively. The quantities in (6) can be estimated using observed data. An alternative standard error estimate can be derived by extending (5.10) in [15]. Using $\hat{\theta}$ and its standard error, we can compare the AUCs for different biomarkers and hence their overall predictive accuracies when biomarkers are measured from independent samples.

Biomarker measurements are often correlated such as the study in our data example. The estimated variance needs to be modified whereas the point estimate stays the same. In case of comparison of correlated biomarkers, we generalized the estimated variance in DeLong et al. [14] to account for the batch effect. More generally, when correlated observations are used to compute the AUC for one biomarker, we generalized the estimates of the variance in [16] to account for the batch effect (Appendix B). Variance can also be estimated by extending bootstrap or jackknife procedures for correlated data [17].

2.4. Implications of Study Designs

Study design plays an important role in the choice of the proposed methods, in particular when Y is an indicator for disease status. First, following Prentice and Pike [18], one can show that case-control studies can be analyzed using the proposed methods as if they were prospective studies. While the interpretation of the intercept term in a logistic regression is no longer valid, the interpretation for γ and θ still hold. Second, the choice of the quantity ζ_b may depend on the design of the study. For a cohort study, ζ_b can be taken as the median in all subjects in the b th batch. On the other hand, for a case-control study, the median of all subjects in the b th batch is difficult to interpret; instead, it may be more appropriate to use the median of only controls (or cases) in the b th batch.

Study design also plays an important role in the performance of the proposed methods, since their performance depends on the size of m_{1b} and m_{0b} . For example, when the prevalence of a disease is low, the number of cases will be low within each batch in a cohort study, but it is fixed under a case-control study. Intuitively, given the total number of observations per

batch, it is preferable to have an equal number of cases and controls to achieve better performance. The impact of study design will be examined in more details in our simulation studies. Our conclusions are consistent with recommendations made in [1, 2, 3].

3. Simulation Studies

We conducted simulation studies to examine finite sample performances of the proposed estimators and factors that may impact their performances. We considered the following simulation settings. Suppose Y is an indicator for disease status, 1 for diseased and 0 for non-diseased, and $\pi(X, V) = E(Y|X, V)$, and Y follows a logit model as follows

$$\text{logit}(\pi(X, V)) = \alpha + \beta_1 X + \beta_2 V \quad (7)$$

where V introduces confounding. X and V were generated independently from a standard normal distribution. As we discussed in Section 2.2, the simulated true model (7) and the distribution of X induce a true model (3).

To assess the marginal association of X and Y , two binary predictors were defined to fit model (3) and (4), $Z_{Tbi} = I(X_{bi} \leq \xi_X)$, where ξ_X is the estimated overall median of X in controls, $Z_{bi} = I(W_{bi} \leq \xi_b)$, where ξ_b is the sample median of W in the controls in the b th batch, and $Z_{1bi} = I(W_{bi} \leq \zeta)$, where ζ is the median of W in all controls. Three methods were considered: a) a method (GS) using the unobserved Z_T to fit model (3); b) the proposed robust method (RM) using Z to fit model (4); and c) a naive method (N) using contaminated Z_1 to fit model (4). Method GS used the true dichotomized values and was considered the gold standard. Similarly, to assess the predictive accuracy of X , three methods were compared for estimating AUC (θ), a method (GS) that computes an usual U statistic, the formula (5.5) in [15], using the unobserved X_1 and X_0 , the proposed robust method (RM) using (5), and a naive method (N) that computes an usual U statistic, the formula (5.5) in [15], using contaminated measurements W_1 and W_0 . We note that the GS method is not applicable in practice, and it only serves as a comparison with our proposed method in simulation studies. The true value of γ and θ are denoted by γ_T and θ_T , which are obtained numerically since an analytical form is not available. Usual model-based standard errors were computed for $\hat{\gamma}$'s, and standard errors based on our formula (6) and the formula (5.12) in [15] were computed for $\hat{\theta}$.

X and V are generated independently from a standard normal distribution. Three error structures were considered: a) additive, $W_{bi} = X_{bi} + \eta_b$; b) multiplicative, $W_{bi} = X_{bi}\eta_b$; and c) power function $W_{bi} = (X_{bi} + 10)^{\eta_b}$. η_b is assumed to follow a uniform distribution on $(0, s)$, where s indicates the maximum magnitude of measurement errors. In addition, the impact of several factors was studied: 1) α and β_1 that represent different baseline prevalence of the disease and different strength of the association between Y and X ; 2) $n_b = m_{1b} + m_{0b}$, the sample size within each batch; and 3) B , the number of batches. The simulations were conducted for both cohort and case-control study designs. It was assumed in all our simulations that the number of observations per batch is the same ($n_b = n_B$ for all b), and for case-control studies there were an equal number of cases and controls in each batch ($m_{1b} = m_{0b}$).

The main results of our simulation studies are summarized in Table I–IV when assuming the power function for measurement errors ($s = 2$). Standard errors of $\hat{\theta}$ in Table I–IV were computed using Equation (6). The results using other error structures are similar. In general, the performance of our proposed RM estimator for both γ and θ is satisfactory under various settings and is comparable to that of the GS method. The naive estimator of γ and θ always

shows substantial bias. For a small sample size of 10 observations per batch, the performance of the proposed methods is still acceptable. As the number of observations per batch (n_B) increases, the performance of our RM estimator improves drastically, whereas the naive estimator still exhibits considerable bias. When the number of batches increases, say from 30 to 150, there are small changes in the proposed RM estimator for γ and very little change in the RM estimator of θ . For the same finite sample size, the proposed RM estimator for AUC achieves better performance compared to that for log odds ratio, γ .

For a cohort study, our simulation studies show that as the association between exposure and outcome becomes stronger ($\beta_1 = 2$), the difference between our RM estimator and GS estimator for γ becomes larger, but the difference for AUC (θ) becomes smaller. As the disease prevalence rate decreases ($\alpha = -2$), both differences become larger. This observation is not surprising, since as the disease prevalence rate decreases, the number of cases per batch would decrease which would in turn affect the performances of our RM estimators, which are stratified on batches. For a case-control study, as the association between exposure and outcome increases, the difference between the GS estimator and RM estimator becomes smaller for both γ and θ . For a case-control study, the number of cases and controls is fixed in each batch, therefore we expected that disease prevalence rate has very little effect on the performances of our RM estimators, which was confirmed by our simulation studies. These observations suggest that a case-control study design is preferred when the prevalence rate of a disease is low.

Our simulation results also show that both the model-based standard error for $\hat{\gamma}$ and the proposed standard error using (6) for $\hat{\theta}$ achieve satisfactory performance, i.e., close to the SD, and they tend to be lower than the SE and SD of the corresponding GS estimator. However, the standard errors extending (5.12) in [15] tend to underestimate the true sampling variation of $\hat{\theta}$ when n_B is small. Additional simulation studies were conducted to study the performance of conditional score and corrected score methods for fitting model (2) under the misspecified error structure, and these estimators exhibited substantial bias in estimating β in (2). Since these methods assume normally distributed errors and an additive structure, our findings are not unexpected.

4. Data Analysis: a Study of Biomarkers of Risk for Colorectal Cancer

We illustrate the proposed methods using the Markers of Adenomatous Polyps II (MAP II) study. The MAP II study was a pilot, biomarker validation case-control study [19, 20]. MAP II recruited adult subjects with no history of previous colorectal adenoma or malignancy of any type, who were scheduled for elective outpatient colonoscopy at a large private practice gastroenterology group. Participants with adenoma were considered “cases” with $Y = 1$, and participants with no adenoma were considered “controls” with $Y = 0$.

A panel of plausible protein biomarkers that describes molecular phenotypes of the normal-appearing colorectal epithelium have been developed [19, 20, 21], which represent highlights of features of the known molecular basis of the earliest stages of colorectal carcinogenesis. Tissue samples from patients were first immunohistochemically processed to identify biomarkers; then using a novel, custom developed image analysis program, biomarkers were measured along colon crypts, microscopic structures in the human colon mucosa. This unique image analysis program is the only currently available approach that can measure protein biomarker at the crypt level. Biomarker distribution curves were constructed from the staining optical density data by standardizing the crypt length into 50 segments (where 1 denotes the first cell at the base and 50 the apical cell of the crypt) and plotting the mean optical density across crypts against the segment location. Due to the limitation of staining process, in general 40 samples from 8 subjects can be processed at the

same time in each batch. Also due to other unforeseen complications, the entire samples were measured at different times and locations by different technicians.

For the present data analysis, the biomarker measurements were averaged over 50 segments of each crypt and only biopsy samples from the rectum were used. We used two biomarkers, TGF- α and Bax, to illustrate our methods for assessing the association between biomarkers and case/control status and comparing the predictive accuracy of multiple biomarkers. Due to the limitation of the current technology, different biomarkers need to be processed and measured separately on different samples, which resulted in different sample sizes. For TGF- α , a total of 31 cases and 31 controls from 10 batches were used in our data analysis, and on average 32.7 crypts were scored for each subject. For Bax, a total of 42 cases and 45 controls from 12 batches were used in our data analysis, and on average 25.4 crypts were scored for each subject. For TGF- α , its measurement ranged from 0 to 1504, and its mean for each batch ranged from 21.7 to 577.5. Large variation between batches and considerably smaller variation within each batch were observed, and these observations indicated potential non-additive measurement errors or non-constant variance for measurement errors. Similar patterns were observed for Bax. Similar to what is discussed in Section 1 and 2.1, the conditions for each batch were held constant but varied from one batch to another in this study; since the measurement errors are primarily the result of the immunohistochemical process, it is reasonable to assume that the measurement error is constant within each batch.

Previously, the data were analyzed [20] using model (2), and that analysis first standardized the biomarker measurements by assuming both an additive measurement structure and constant measurement error within each batch. This data set was reanalyzed using methods for models (4) and (5) without the assumption of additive error structure. For the purpose of comparison, the naive method was also used to analyze this data set.

For the ROC curve analysis, we considered each crypt as one observation. In addition to compute the AUC values, we also compared the standard errors that accounts for correlation and that assume independence, respectively. The results are summarized in Table V. Our results show that without adjusting for batch effect and using the naive estimator (N), the estimates of the AUC for both biomarkers are biased towards the null ($\theta = 0.5$). Using our proposed robust method (RM), the estimate of the AUC increased dramatically to 0.615 and 0.676 for Bax and TGF- α , respectively, and their differences with 0.5 are statistically significant with $p < 0.001$. It also appears that TGF- α has better diagnostic accuracy than Bax, and the difference in AUC between two biomarkers increases to 0.061. Our results indicate that the diagnostic accuracy of these two biomarkers are fair, though not excellent. In addition, our results show that standard errors are underestimated if one does not adjust for correlations between multiple measurements per subject.

For the association analysis, the average staining optical density was taken over all crypts for each subject. The same transformation in our simulations was used, that is, the median of the controls in each batch was used to define Z and the median in all controls to define Z_1 for comparisons. A logistic regression was then fitted and the odds ratios were calculated using Z and Z_1 , respectively. Table VI summarizes our results. Our results indicate that measurement error biased the estimates of the OR towards the null (OR=1). In particular, for the biomarker TGF- α , the proposed analysis shows that the cases were more likely to have higher TGF- α (OR=3.01) and the association was statistically significant ($p = 0.04$), whereas the analysis without adjustment for measurement errors did not reveal a strong association (OR=1.48) and it was not statistically significant ($p = 0.45$). For biomarker Bax, two methods also led to different point estimates of OR, though the confidence intervals for both OR estimates include 1.

In summary, our analyses indicate that there were relatively strong association between the expression of biomarker TGF- α and the presence of adenoma, but the association between the expression of Bax and the presence of adenoma is not statistically significant.

5. Discussion

In this paper, we propose two robust statistical methods to analyze data in the presence of batch/experiment specific errors. These methods require no assumptions on the structure and distribution of the measurement error. Compared to methods predicated on normality and additive structure of measurement errors, our methods yield valid inferences under departure from these assumptions. In practice, when these structural and distributional assumptions are questionable, some have suggested to attempt certain transformations so that these assumptions become reasonable for transformed data. The task of finding such transformations can be challenging, and our proposed methods provide a viable alternative. The proposed methods are easy to implement using off-shelf software after simple variable transformations, and their performance is satisfactory even for a relatively small sample size. The design of a study has important implications. In the presence of such batch/experiment specific errors, investigators may want to assign equal numbers of cases and controls to the same batch to improve the performance of the data analysis.

The proposed methods depend on the assumption of batch/experiment specific errors. This assumption is plausible in many experiment settings as well as for some meta-data analysis, when the measurement error or variation is primarily due to varying lab/experiment conditions that can be held steady in each batch/experiment but is hard to replicate across different batch/experiments. When this assumption is in question, investigators need to proceed with caution and some sensitivity analysis can be conducted, which is a future research topic.

Acknowledgments

This work was supported by a grant from National Cancer Institute, National Institutes of Health (CA114456).

References

1. Blanck HM, Bowman BA, Cooper GR, Myers GL, Miller DT. Laboratory issues: use of nutritional biomarkers. *The Journal of Nutrition*. 2003; 133 (Suppl 3):888S–894S. [PubMed: 12612175]
2. Rundle AG, Vineis P, Ahsan H. Design options for molecular epidemiology research within cohort studies. *Cancer Epidemiology Biomarkers & Prevention*. 2005; 14:1899–1907.
3. Tworoger SS, Yasui Y, Chang L, Stanczyk FZ, McTiernan A. Specimen allocation in longitudinal biomarker studies: controlling subject-specific effects by design. *Cancer Epidemiology Biomarkers & Prevention*. 2004; 13:1257–1260.
4. 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. *American Journal of Epidemiology*. 2009; 170:88–94. [PubMed: 19406919]
5. Ye H, Yu T, Temam S, Ziober BL, Wang J, Schwartz JL, Mao L, Wong DT, Zhou X. Transcriptomic dissection of tongue squamous cell carcinoma. *BMC Genomics*. 2008; 9:69. [PubMed: 18254958]
6. Yu T, Ye H, Chen Z, Ziober BL, Zhou X. Dimension reduction and mixed-effects model for microarray meta-analysis of cancer. *Frontiers in Bioscience*. 2008; 13:2714–2720. [PubMed: 17981746]
7. Carroll, RJ.; Ruppert, D.; Stefanski, LA.; Crainiceanu, CM. Measurement error in nonlinear models: a modern Perspective. 2nd Edition. Chapman & Hall/CRC; New York: 2006.

8. Stefanski LA, Carroll RJ. Conditional scores and optimal scores in generalized linear measurement error models. *Biometrika*. 1987; 74:703–716.
9. Nakaruma T. Corrected score functions for errors-in-variables models: methodology and application to generalized linear models. *Biometrika*. 1990; 77:127–137.
10. Stefanski LA. Unbiased estimation of a nonlinear function of a normal mean with application to measurement error models. *Communications in Statistics Theory and Methods*. 1989; 18:4335–4358.
11. Stefanski LA, Buzas JS. Instrumental variable estimation in binary regression measurement error models. *Journal of the American Statistical Association*. 1995; 90:541–550.
12. Cook JR, Stefanski LA. Simulation-extrapolation estimation in parametric measurement error models. *Journal of the American statistical Association*. 1994; 89:1314–1328.
13. Bamber D. The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. *Journal of Mathematical Psychology*. 1975; 12:387–415.
14. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*. 1988; 44:837–845. [PubMed: 3203132]
15. Pepe, MS. *The Statistical Evaluation of Medical Tests for Classification and Prediction*. Oxford University Press; Oxford, United Kingdom: 2003.
16. Obuchowski NA. Nonparametric analysis of clustered ROC curve data. *Biometrics*. 1997; 53:567–578. [PubMed: 9192452]
17. Beam CA. Analysis of clustered data in receiver operating characteristic studies. *Stat Methods Med Res*. 1998; 7:324–336. [PubMed: 9871950]
18. Prentice RL, Pyke R. Logistic disease incidence models and case-control studies. *Biometrika*. 1979; 66:403–411.
19. Ahearn T, Dash C, Bostick R. Associations of calcium and vitamin D with E-cadherin and β -catenin expression in normal-appearing rectal tissue; markers of adenomatous polyp II (MAPII) case-control study. *Proc AACR*. 2008; 49:133.
20. 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 & Prevention*. 2009; 18(1):65–73.
21. Fedirko V, Bostick RM, Flanders WD, Long Q, Shaikat A, Rutherford RE, Daniel CR, Cohen V, Dash C, Woodard JJ. Effects of vitamin D and calcium supplementation on markers of apoptosis in normal colon mucosa: A randomized, double-blind, placebo-controlled clinical trial. *Cancer Prevention Research*. 2009; 2(3):213–223. [PubMed: 19258546]
22. Maritz JS, Jarrett RG. A note on estimating the variance of the sample median. *Journal of the American Statistical Association*. 1978; 73(361):194–196.
23. Sukhatme S, Beam CA. Stratification in nonparametric ROC studies. *Biometrics*. 1994; 50:149–163. [PubMed: 8086598]
24. Hoeffding W. A class of statistics with asymptotically normal distribution. *Annals of Mathematical Statistics*. 1978; 19:293–325.

APPENDIX

A: Asymptotic Properties of $\hat{y}$

For simplicity, we assume ξ_X is the true median of X , and the proof can be extended to other order statistics. Following the notation in Section 2.1, $Z_{bi} = I(X_{bi} \leq \xi_X)$, and $Z_{bi}^* = I(W_{bi} \leq \hat{\xi}_b)$, where ξ_b and $\hat{\xi}_b$ is the median of W in the b th batch and its sample estimate. Z_{bi} is not observed whereas Z_{bi}^* is.

It is well known [22] that

$$\sqrt{n_b}(\widehat{\xi}_b - \xi_b) \rightarrow_d N(0, \frac{1}{4f_b^2(\xi_b)}).$$

where $f_b(w)$ is the density function of W_b in the b th batch evaluated at w . Then it follows that $I(W_{bi} \leq \widehat{\xi}_b)$ converges to $I(W_{bi} \leq \xi_b)$ in probability. When the rank-preserving assumption holds, it is straightforward to show that $f_b(\xi_b) = f(\xi_X)$ with f being the density function of X and $I(W_{bi} \leq \xi_b) = I(X_{bi} \leq \xi_X)$ since $W_{bi} = g(X_{bi}, \eta_b)$ and $\xi_b = g(\xi_X, \eta_b)$; and hence Z_{bi}^* converges to Z_{bi} in probability as $n_b \rightarrow \infty$.

For simplicity, we consider a setting where the canonical link is used for model (2), and our results hold for other link functions. As a result of fitting model (3) and (4), some algebra can show that

$$\begin{aligned}\widehat{\gamma}_T &= g\left(\frac{\sum_b \sum_i Z_{bi} Y_{bi}}{\sum_b \sum_i Z_{bi}}\right) - g\left(\frac{\sum_b \sum_i (1-Z_{bi}) Y_{bi}}{\sum_b \sum_i (1-Z_{bi})}\right) \\ \widehat{\gamma} &= g\left(\frac{\sum_b \sum_i Z_{bi}^* Y_{bi}}{\sum_b \sum_i Z_{bi}^*}\right) - g\left(\frac{\sum_b \sum_i (1-Z_{bi}^*) Y_{bi}}{\sum_b \sum_i (1-Z_{bi}^*)}\right)\end{aligned}$$

Under some usual regularity conditions, due to $Z_{bi}^* - Z_{bi} \rightarrow_p 0$, one can show that

$\sum_b \sum_i Z_{bi}^*/n - \sum_b \sum_i Z_{bi}/n \rightarrow_p 0$ and $\sum_b \sum_i Z_{bi}^* Y_{bi}/n - \sum_b \sum_i Z_{bi} Y_{bi}/n \rightarrow_p 0$. It follows that $\widehat{\gamma} - \widehat{\gamma}_T \rightarrow_p 0$ as $n_b \rightarrow \infty$; and since $\widehat{\gamma}_T \rightarrow_p \gamma_T$ as $n_b \rightarrow \infty$, we have $\widehat{\gamma} \rightarrow_p \gamma_T$ as $n_b \rightarrow \infty$.

B: Properties of $\hat{\theta}$ for a ROC Analysis

Using observed data W and Y , the proposed estimator of θ is $\hat{\theta} = \hat{\Sigma}_b w_b \theta_b$, where $w_b = (m_{1b} + m_{0b})/(m_1 + m_0)$ is a weight function and $\hat{\theta}_b$ is the Mann-Whitney U statistics within the b th batch and equal to $\frac{1}{m_{1b}m_{0b}} \sum_{i=1}^{m_{1b}} \sum_{j=1}^{m_{0b}} \psi(W_{1bi}, W_{0bj})$.

Following similar arguments in [23], one can show that

$$\widehat{\theta}_T = \sum_{b=1}^B w_b \frac{1}{m_{1b}m_{0b}} \sum_{i=1}^{m_{1b}} \sum_{j=1}^{m_{0b}} \psi(X_{1bi}, X_{0bj})$$

is an unbiased estimator of the AUC. We note that $\widehat{\theta}_T$ is based on the unobserved data X_1 and X_0 , and can not be used to estimate θ in practice. If the rank preserving property holds, we have $\psi(W_{1bi}, W_{0bj}) = \psi(X_{1bi}, X_{0bj})$ for all pairs of i and j . It is straightforward to show that $\hat{\theta}$ is equal to $\widehat{\theta}_T$ and hence is an unbiased estimator of the AUC.

Within each batch, one can show

$$\widehat{Var}(\widehat{\theta}_b) = \frac{1}{m_{1b}m_{0b}} \{ \theta_b(1 - \theta_b) + (m_{1b} - 1)(R_{1b} - \theta_b^2) + (m_{0b} - 1)(R_{0b} - \theta_b^2) \}.$$

It follows that

$$\widehat{Var}(\widehat{\theta}) = \sum_{b=1}^B \frac{w_b^2}{m_{1b}m_{0b}} \{ \theta_b(1 - \theta_b) + (m_{1b} - 1)(R_{1b} - \theta_b^2) + (m_{0b} - 1)(R_{0b} - \theta_b^2) \}$$

Applying the theory of U-statistics in [24], one can show that

$$\sqrt{m_1 + m_0}(\widehat{\theta} - \theta) / \sqrt{\widehat{Var}(\widehat{\theta})}$$

is asymptotically normally distributed with mean 0 and variance 1 as $m_1 + m_0 \rightarrow \infty$.

C: Properties of $\widehat{\theta}$ for Correlated Data

Let X_{1bij} and X_{0bij} denote the j th ($j = 1, \dots, m_{1bi}$ or $j = 1, \dots, m_{0bi}$) observed biomarker value for the i th cluster ($i = 1, \dots, I_b$) in the b th batch for cases and controls, respectively. We assume that clusters are nested within batches. Let $m_{1b} = \sum_i m_{1bi}$ and $m_{0b} = \sum_i m_{0bi}$. The estimate of the AUC is

$$\widehat{\theta} = \sum_b w_b \widehat{\theta}_b$$

where $\widehat{\theta}_b = \frac{1}{m_{1b}m_{0b}} \sum_{i=1}^{I_b} \sum_{i'=1}^{I_b} \sum_{j=1}^{m_{1bi}} \sum_{k=1}^{m_{0bi}} \psi(W_{1bij}, W_{0bi'k})$. Then an estimator of the variance of $\widehat{\theta}$ is

$$\widehat{var}(\widehat{\theta}) = \sum_{b=1}^B w_b^2 \widehat{var}(\widehat{\theta}_b)$$

Similar to formula (4) in [16], we can obtain $\widehat{var}(\widehat{\theta}_b)$ and hence $\widehat{var}(\widehat{\theta})$. Similar to formula (5) in [16], we can compute the estimated covariance between $\widehat{\theta}$'s for two biomarkers that are measured on the same units.

Table 1

Simulation results for a cohort design: the impact of α and β_1 with $n_B = 10$, $B = 30$, $\beta_2 = 0.5$, $W = (X + 10)\eta$ and η following uniform $(0, 2)$; EST, the estimate; SE, average of estimated standard errors; SD, Monte Carlo standard deviation. Estimators: GS, the method using the unobserved X ; RM, the proposed robust method; N, the naive method.

Setting	True Value	Method	EST	SE	SD
$\alpha = 0, \beta_1 = 1$	$\gamma_T = 1.433$	GS	1.456	0.267	0.276
		RM	1.460	0.258	0.258
		N	0.099	0.232	0.233
$\alpha = -2, \beta_1 = 1$	$\gamma_T = 1.513$	GS	1.577	0.411	0.427
		RM	1.556	0.395	0.403
		N	0.098	0.318	0.325
$\alpha = -2, \beta_1 = 2$	$\gamma_T = 2.830$	GS	2.968	0.605	0.621
		RM	2.749	0.520	0.546
		N	0.151	0.279	0.291
$\alpha = 0, \beta_1 = 1$	$\theta_T = 0.730$	GS	0.731	0.029	0.027
		RM	0.754	0.023	0.025
		N	0.532	0.031	0.024
$\alpha = -2, \beta_1 = 1$	$\theta_T = 0.740$	GS	0.740	0.038	0.037
		RM	0.791	0.032	0.031
		N	0.542	0.042	0.032
$\alpha = -2, \beta_1 = 2$	$\theta_T = 0.863$	GS	0.863	0.024	0.024
		RM	0.872	0.026	0.024
		N	0.544	0.039	0.032

Simulation results for a cohort design: the impact of n_B and B with $W = (X + 10)^{\eta}$ and η following uniform (0, 2); EST, the estimate; SE, average of estimated standard errors; SD, Monte Carlo standard deviation. Estimators: GS, the method using the unobserved X ; RM, the proposed robust method; N, the naive method.

Table II

Setting	True Value	Method	EST	SE	SD
$\alpha = -2, \beta_1 = 2, \beta_2 = 0.5$ for estimating γ					
$n_B = 10, B = 30$	$\gamma_T = 2.830$	GS	2.968	0.605	0.621
		RM	2.749	0.520	0.546
		N	0.151	0.279	0.291
$n_B = 20, B = 30$		GS	2.893	0.405	0.430
		RM	2.777	0.374	0.400
		N	0.149	0.196	0.204
$n_B = 50, B = 30$		GS	2.869	0.249	0.262
		RM	2.820	0.241	0.244
		N	0.159	0.124	0.139
$n_B = 10, B = 150$		GS	2.853	0.247	0.271
		RM	2.634	0.214	0.223
		N	0.112	0.123	0.129
$\alpha = -2, \beta_1 = 1, \beta_2 = 0.5$ for estimating θ					
$n_B = 10, B = 30$	$\theta_T = 0.740$	GS	0.740	0.038	0.037
		RM	0.791	0.032	0.031
		N	0.542	0.042	0.032
$n_B = 20, B = 30$		GS	0.740	0.027	0.027
		RM	0.761	0.022	0.026
		N	0.533	0.031	0.024
$n_B = 50, B = 30$		GS	0.740	0.017	0.016
		RM	0.741	0.018	0.017
		N	0.528	0.020	0.018
$n_B = 10, B = 150$		GS	0.739	0.017	0.018
		RM	0.790	0.010	0.013
		N			

Setting	True Value	Method	EST	SE	SD
		N	0.523	0.020	0.016

Simulation results for a case-control design: the impact of α and β_1 with $n_B = 10$, $B = 30$, $\beta_2 = 0.5$, $W = (X + 10)^\eta$ and η following uniform (0, 2); EST, the estimate; SE, average of estimated standard errors; SD, Monte Carlo standard deviation. Estimators: GS, the method using the unobserved X ; RM, the proposed robust method; N, the naive method.

Table III

Setting	True Value	Method	EST	SE	SD
$\alpha = 0, \beta_1 = 1$	$\gamma_T = 1.433$	GS	1.446	0.266	0.273
		RM	1.673	0.259	0.244
		N	0.094	0.231	0.074
$\alpha = -2, \beta_1 = 1$	$\gamma_T = 1.513$	GS	1.538	0.271	0.293
		RM	1.752	0.263	0.256
		N	0.099	0.231	0.077
$\alpha = -2, \beta_1 = 2$	$\gamma_T = 2.830$	GS	2.900	0.412	0.449
		RM	2.857	0.351	0.385
		N	0.145	0.231	0.088
$\alpha = 0, \beta_1 = 1$	$\theta_T = 0.730$	GS	0.730	0.028	0.027
		RM	0.748	0.023	0.024
		N	0.524	0.033	0.005
$\alpha = -2, \beta_1 = 1$	$\theta_T = 0.740$	GS	0.740	0.028	0.029
		RM	0.756	0.024	0.026
		N	0.525	0.033	0.005
$\alpha = -2, \beta_1 = 2$	$\theta_T = 0.863$	GS	0.863	0.020	0.021
		RM	0.865	0.021	0.022
		N	0.537	0.033	0.007

Simulation results for a case-control design: the impact of n_B and B with $\alpha = -2$, $\beta_1 = 1$, $\beta_2 = 0.5$, $W = (X + 10)^\eta$ and η following uniform $(0, 2)$; EST, the estimate; SE, average of estimated standard errors; SD, Monte Carlo standard deviation. Estimators: GS, the method using the unobserved X ; RM, the proposed robust method; N, the naive method.

Table IV

Setting	True Value	Method	EST	SE	SD
$n_B = 10, B = 30$	$\eta_T = 1.513$	GS	1.538	0.271	0.293
		RM	1.752	0.263	0.256
		N	0.099	0.231	0.077
$n_B = 20, B = 30$		GS	1.526	0.190	0.202
		RM	1.424	0.187	0.189
		N	0.098	0.163	0.067
$n_B = 50, B = 30$		GS	1.517	0.120	0.129
		RM	1.545	0.119	0.126
		N	0.099	0.103	0.059
$n_B = 10, B = 150$		GS	1.518	0.120	0.132
		RM	1.731	0.117	0.113
		N	0.079	0.103	0.032
$n_B = 10, B = 30$	$\theta_T = 0.740$	GS	0.740	0.028	0.029
		RM	0.756	0.024	0.026
		N	0.525	0.033	0.005
$n_B = 20, B = 30$		GS	0.739	0.020	0.020
		RM	0.741	0.020	0.020
		N	0.524	0.023	0.005
$n_B = 50, B = 30$		GS	0.739	0.013	0.013
		RM	0.739	0.013	0.013
		N	0.525	0.015	0.004
$n_B = 10, B = 150$		GS	0.739	0.013	0.013
		RM	0.756	0.011	0.012
		N	0.519	0.015	0.002

Table V

Estimated AUC for ROC using the MAP II data: RM, the proposed robust method; N, the naive method.

Estimate of AUC		
Biomarker	RM	N
TGF- α	0.676	0.559
SE ₁ ^a	0.039	0.066
SE ₂ ^b	0.006	0.013
Bax	0.615	0.516
SE ₁	0.028	0.064
SE ₂	0.005	0.013

^a standard error accounting for the correlations.

^b standard error assuming independence.

Table VI

Estimated Odds Ratio using the MAP II data: RM, the proposed robust method; N, the naive method.

Biomarker	Estimate of OR	
	<i>RM</i>	<i>N</i>
TGF- α	3.01	1.48
CI ^a	(1.07,8.45)	(0.55,3.96)
Bax	0.88	0.74
CI	(0.37,2.07)	(0.32,1.73)

^a95% confidence interval.