Special Issue Paper

Statistics in Medicine

Received 14 January 2011, Accepted 5 September 2011 Published online 13 March 2012 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.4426

Likelihood-based methods for regression analysis with binary exposure status assessed by pooling

Robert H. Lyles,[a*†] Li Tang,[a] Ji Lin,[a] Zhiwei Zhang[b] and Bhramar Mukherjee[c]

The need for resource-intensive laboratory assays to assess exposures in many epidemiologic studies provides ample motivation to consider study designs that incorporate pooled samples. In this paper, we consider the case in which specimens are combined for the purpose of determining the presence or absence of a pool-wise exposure, in lieu of assessing the actual binary exposure status for each member of the pool. We presume a primary logistic regression model for an observed binary outcome, together with a secondary regression model for exposure. We facilitate maximum likelihood analysis by complete enumeration of the possible implications of a positive pool, and we discuss the applicability of this approach under both cross-sectional and case-control sampling. We also provide a maximum likelihood approach for longitudinal or repeated measures studies where the binary outcome and exposure are assessed on multiple occasions and within-subject pooling is conducted for exposure assessment. Simulation studies illustrate the performance of the proposed approaches along with their computational feasibility using widely available software. We apply the methods to investigate gene–disease association in a population-based case-control study of colorectal cancer. Copyright © 2012 John Wiley & Sons, Ltd.

Keywords: cross-sectional study; case-control study; efficiency; logistic regression; pooling; repeated measures; single nucleotide polymorphism

1. Introduction

Laboratory assays to assess the presence and/or level of a biomarker (or, more generically, an exposure) are often a critical component of epidemiologic research. Unfortunately, such assays can be expensive and time-consuming, posing significant limitations upon the number of subjects that can feasibly be included in a particular study. As a result, the concept of combining samples to reduce the overall numbers of assays performed has frequently been discussed in the literature. Statisticians refer to this process by various names, including Dorfman sampling, group testing, and composite sampling [1–4]. Herein, we elect to use the recently popular term, ‘pooling’ [5–13], which encompasses variations on this theme that include detecting the presence or absence of a disease or substance and measuring the amount present. In the latter case, some authors emphasize potential benefits of pooling for reducing loss of information from assay nondetectables [5, 9], in addition to its appeal in terms of reducing laboratory costs.

In this report, we focus upon pooling as a means of conserving resources when the laboratory assay in question aims at determining the presence or absence of a biomarker or exposure. We assume that the objective of the epidemiological study is to associate exposure status at the individual level (X) to a health-related outcome (Y), controlling for subject-specific covariates ($\boldsymbol{C}$). Although we assume binary

[a] *Department of Biostatistics and Bioinformatics, The Rollins School of Public Health, Emory University, 1518 Clifton Rd. N.E., Atlanta, GA 30322, USA*

[b] *Biostatistics and Bioinformatics Branch, Division of Epidemiology, Statistics, and Prevention Research, Eunice Kennedy Shriver National Institute of Child Health and Human Development, Bethesda, MD 20892, USA*

[c] *Department of Biostatistics, School of Public Health, University of Michigan, Ann Arbor, MI 48109, USA*

**Correspondence to: Robert Lyles, Department of Biostatistics and Bioinformatics, The Rollins School of Public Health, Emory University, 1518 Clifton Rd. N.E., Atlanta, GA 30322, USA.*

[†] *E-mail: rlyles@sph.emory.edu*

outcomes and work with logistic regression models throughout, the methods illustrated are not limited to that setting.

Our work has natural connections with prior demonstrations of pooled data analysis, in which pooling was advocated as a means of assessing the binary outcome (rather than exposure) in univariate or multivariate logistic regression modeling [7, 14]. We note that the issue of potential errors in accurate assay quantification (e.g., average allele frequency) can be viewed as one of the benefits of dichotomizing exposure. That is, it should be easier to make the binary determination of allele presence or absence in a pooled sample. With this in mind, our approach assumes only that the correct yes/no decision about exposure is made for each pool.

In Section 2, we formulate the problem of interest and provide a general maximum likelihood (ML) solution in the univariate case that is readily applicable for pooled and 'hybrid' pooled–unpooled designs [12]. We then show how the ML approach extends naturally to the repeated measures setting via a generalized linear mixed model. Section 3 discusses practical issues including applicability of the univariate approach under a case-control design, and provides some general guidelines regarding efficient pooling strategies. Section 4 contains an example based on associating the presence/absence of a risk allele at a single nucleotide polymorphism (SNP) with disease status in a case-control study of colorectal cancer, and Section 5 presents simulation studies to assess the performance of the methods. We conclude with a discussion in Section 6.

2. Methods

2.1. Model specification and maximum likelihood in the univariate case

To begin, we assume interest lies in the coefficients and associated adjusted odds ratios (ORs) based on the following logistic regression model:

$$\text{logit}[\Pr(Y = 1|X, \boldsymbol{C})] = \alpha + \beta x + \sum_{t=1}^{\mathrm{T}} \gamma_t c_{t\geq}\ (x = 0, 1), \qquad (1)$$

where X represents a binary exposure of interest (e.g., presence/absence of a biomarker) and $\boldsymbol{C} = (C_1, \ldots, C_T)$ is an arbitrary set of covariates which could include, for example, interaction terms that involve X. Our focus is on the scenario in which some or all of the individual subject-specific exposure indicators (X_i, $i = 1, \ldots, n$) are measured only indirectly, after pooling. Commonly, this means that biological samples from two or more subjects are combined and analyzed via a single laboratory assay. We assume that the assay correctly determines the presence or absence of the signal (e.g., biomarker) in the pooled sample, and that this determination translates into the knowledge that either (a) zero or (b) one or more subjects in the pool are positive for exposure.

If one seeks only to estimate the OR (e^{β}) associated with X, an expedient approach would be to treat X as the outcome, that is, via the model

$$\text{logit}[\Pr(X = 1|Y, \boldsymbol{C})] = \alpha^* + \beta y + \sum_{t=1}^{\mathrm{T}} \gamma_t^* c_t \quad (x = 0, 1). \qquad (2)$$

Model (2) targets the same underlying adjusted OR associating Y with X as does model (1), but different adjusted ORs corresponding to the C_js. Technically, if (1) holds exactly then (2) may require higher-order terms in the C_js to be completely valid [15, 16]. Nevertheless, the model fitting exercise for (2) is no more demanding than that for (1). Existing methods [7] for dealing with pooled binary outcome data could be directly applied for the ML estimation of β based on (2).

More commonly, one is interested in the entire set of regression parameters ($\beta, \boldsymbol{\gamma}$) under (1). To set the stage for the proposed approach when pooling of samples is used to determine exposure (X) status, assume a second model that regresses X on $\boldsymbol{C}^*$, where $\boldsymbol{C}^*$ is a vector of covariates that may include some or all of the Cs in model (1) and/or additional variables:

$$\text{logit}[\Pr(X = 1|\boldsymbol{C}^*)] = \psi_0 + \sum_{s=1}^{\mathrm{S}} \psi_s c_s^*. \qquad (3)$$

We specify the logit link in (1) and (3), but note that other links (e.g., probit) could just as easily be used in either model. While it would be completely unnecessary in the absence of pooling if one cares solely about the model (1) parameters, we could consider analyzing the data based on the joint model for (Y, X) given $(\boldsymbol{C}, \boldsymbol{C}^*)$. That is, rather than $\Pr(Y|X, \boldsymbol{C})$, one could work with the following joint likelihood contributions:

$$\Pr(Y, X|\boldsymbol{C}, \boldsymbol{C}^*) = \Pr(Y|X, \boldsymbol{C}) \times \Pr(X|\boldsymbol{C}^*). \tag{4}$$

Assuming distinct sets of parameters for the two component models, the joint likelihood factors into separate pieces as dictated by (1) and (3). Note that, even under misspecification of the model for $X|\boldsymbol{C}^*$ in (3), the maximum likelihood estimate (MLE) for $(\alpha, \beta, \boldsymbol{\gamma})$ based on (4) will be identical to the MLE from the standard logistic model in (1) if there is no pooling.

The joint model in (4) facilitates construction of an appropriate likelihood to handle pooled samples in the context of (1). To clarify the idea, consider the case in which a pooled sample involving only two subjects (i and i) is used to assess X. If the sample is negative, we know that both subjects are negative and the joint likelihood contribution for the pool is given by

$$\begin{aligned} p_{\boldsymbol{y}\boldsymbol{0}} &= \Pr[(Y_i = y_i, Y_{i'} = y_{i'}) \cap (X_i = 0, X_{i'} = 0)|\boldsymbol{C}_i = \boldsymbol{c}_i, \boldsymbol{C}_i^* = \boldsymbol{c}_i^*, \boldsymbol{C}_{i'} = \boldsymbol{c}_{i'}, \boldsymbol{C}_{i'}^* = \boldsymbol{c}_{i'}^*] \\ &= [\Pr(Y_i = y_i|X_i = 0, \boldsymbol{C}_i = \boldsymbol{c}_i) \times \Pr(X_i = 0|\boldsymbol{C}_i^* = \boldsymbol{c}_i^*)] \\ &\quad \times [\Pr(Y_{i'} = y_{i'}|X_{i'} = 0, \boldsymbol{C}_{i'} = \boldsymbol{c}_{i'}) \times \Pr(X_{i'} = 0|\boldsymbol{C}_{i'}^* = \boldsymbol{c}_{i'}^*)], \end{aligned} \tag{5}$$

where each term follows directly from models (1) and (3). For a positive pooled sample, we enumerate the possibilities with respect to X, as follows:

$$\begin{aligned} &\Pr\{(Y_i, Y_{i'}) \cap [(X_i = 1, X_{i'} = 1) \text{ OR } (X_i = 1, X_{i'} = 0) \text{ OR } (X_i = 0, X_{i'} = 1)]|\boldsymbol{C}_i, \boldsymbol{C}_{i'}, \boldsymbol{C}_i^*, \boldsymbol{C}_{i'}^*\} \\ &\quad = \Pr[(Y_i, Y_{i'}), X_i = 1, X_{i'} = 1|\boldsymbol{C}_i, \boldsymbol{C}_{i'}, \boldsymbol{C}_i^*, \boldsymbol{C}_{i'}^*] \\ &\quad\quad + \Pr[(Y_i, Y_{i'}), X_i = 1, X_{i'} = 0|\boldsymbol{C}_i, \boldsymbol{C}_{i'}, \boldsymbol{C}_i^*, \boldsymbol{C}_{i'}^*] \\ &\quad\quad + \Pr[(Y_i, Y_{i'}), X_i = 0, X_{i'} = 1|\boldsymbol{C}_i, \boldsymbol{C}_{i'}, \boldsymbol{C}_i^*, \boldsymbol{C}_{i'}^*] \\ &\quad = \Pr\,(Y_i = y_i|X_i = 1, \boldsymbol{C}_i = \boldsymbol{c}_i) \times \Pr(X_i = 1|\boldsymbol{C}_i^* = c_i^*) \\ &\quad\quad \times \Pr(Y_{i'} = y_{i'}|X_{i'} = 1, \boldsymbol{C}_{i'} = \boldsymbol{c}_{i'}) \times \Pr(X_{i'} = 1|\boldsymbol{C}_{i'}^* = \boldsymbol{c}_{i'}^*) \\ &\quad\quad + \Pr(Y_i = y_i|X_i = 1, \boldsymbol{C}_i = \boldsymbol{c}_i) \times \Pr(X_i = 1|\boldsymbol{C}_i^* = \boldsymbol{c}_i^*) \\ &\quad\quad \times \Pr(Y_{i'} = y_{i'}|X_{i'} = 0, \boldsymbol{C}_{i'} = \boldsymbol{c}_{i'}) \times \Pr(X_{i'} = 0|\boldsymbol{C}_{i'}^* = \boldsymbol{c}_{i'}^*) \\ &\quad\quad + \Pr(Y_i = y_i|X_i = 0, \boldsymbol{C}_i = \boldsymbol{c}_i) \times \Pr(X_i = 0|\boldsymbol{C}_i^* = \boldsymbol{c}_i^*) \\ &\quad\quad \times \Pr(Y_{i'} = y_{i'}|X_{i'} = 1, \boldsymbol{C}_{i'} = \boldsymbol{c}_{i'}) \times \Pr(X_{i'} = 1|\boldsymbol{C}_{i'}^* = \boldsymbol{c}_{i'}^*) \end{aligned} \tag{6}$$

The overall likelihood function is the product of the contributions for the negative [Equation (5)] and the positive pools [Equation (6)].

More generally, suppose we have a pool of arbitrary size (r), and again let $p_{\boldsymbol{y}\boldsymbol{0}}$ represent the probability that the pool is negative. It follows that

$$p_{\boldsymbol{y}\boldsymbol{0}} = \left\{\prod_{i=1}^{r} \Pr(Y_i = y_i|X_i = 0, \boldsymbol{C}_i = \boldsymbol{c}_i)\right\} \times \left\{\prod_{i=1}^{r} \Pr(X_i = 0|\boldsymbol{C}_i^* = \boldsymbol{c}_i^*)\right\}, \tag{7}$$

where the subscript 'i' is used here to indicate the subject within the pool. Next, let $p_{\boldsymbol{y}}$ represent the probability of the observed outcomes for the pooled subjects conditional on the covariates $(\boldsymbol{C}, \boldsymbol{C}^*)$, that is, $p_{\boldsymbol{y}} = \Pr\left\{Y_1 = y_1, ..., Y_r = y_r|\boldsymbol{C}_1 = \boldsymbol{c}_1, \boldsymbol{C}_1^* = \boldsymbol{c}_1^*, ..., \boldsymbol{C}_r = \boldsymbol{c}_r, \boldsymbol{C}_r^* = \boldsymbol{c}_r^*\right)$. Using the law of total probability to sum over the possible values of X for members of the pool, we have

$$p_{\boldsymbol{y}} = \sum_{x_1=0}^{1} \sum_{x_2=0}^{1} \cdots \sum_{x_r=0}^{1} \left\{\prod_{i=1}^{r} \Pr(Y_i = y_i|X_i = x_i, \boldsymbol{C}_i = \boldsymbol{c}_i)\right\} \times \left\{\prod_{i=1}^{r} \Pr(X_i = x_i|\boldsymbol{C}_i^* = \boldsymbol{c}_i^*)\right\}. \tag{8}$$

It follows that the likelihood contribution for a positive pool is $(p_{\boldsymbol{y}} - p_{\boldsymbol{y}\boldsymbol{0}})$, the difference between the probabilities given in (8) and (7).

Note that the joint model in (4) facilitates a maximum likelihood approach to handle pooled binary exposures that is analogous to the strategy proposed by Vansteelandt *et al.* [7] for dealing with pooled binary outcome data. The primary difference here is that complete enumeration is required in (8) because in this case the probabilities of positivity and negativity for a particular pool sum up to p_y rather than to one. Maximization of the likelihood function and estimation of the associated Hessian stemming from the contributions in (7) and (8) is quite feasible using a built-in Quasi-Newton routine available in SAS IML (SAS Institute, Inc., Cary, NC, USA) [17].

We find computation of the likelihood to be noticeably more time-consuming when pools are large (e.g., for pools of size 8 or more), but note that there is no conceptual difficulty with handling pools of varying sizes. As a special case, the approach readily accommodates 'hybrid' designs [12] in which some subjects have X measured individually while the rest are pooled. Those with X measured individually are allocated the basic likelihood contribution in (4).

2.2. *Extension to repeated measures or longitudinal studies*

In this section, we assume that the binary outcome and exposure variables are measured repeatedly or longitudinally, along with other covariates that may or may not be time-dependent. In the absence of pooling, the use of nonlinear mixed models is one common approach to the analysis of such data [18, 19]. To illustrate the extension of the approach in the previous section, consider the following logistic-normal model:

$$\text{logit}[\Pr(Y_{ij}=1|X_{ij},\boldsymbol{C}_{ij},u_{yi})]=\alpha+\beta x_{ij}+\sum_{t=1}^{T}\gamma_t c_{ijt}+u_{yi}. \tag{9}$$

Here $y_{ij}(=0,1)$ and $x_{ij}(=0,1)$ are the values of the repeated outcome and exposure variables of interest and c_{ijt} is the value of the t-th (of T) covariates, each measured on subject i at time point j $(i=1,\ldots,k; j=1,\ldots,n_i)$. We assume the random effects u_{yi} are distributed as $N(0,\sigma^2_{uy})$.

Model (9) is a natural extension of model (1) to incorporate random intercepts. Similarly, consider the following logistic-normal extension of model (3)

$$\text{logit}[\Pr(X_{ij}=1|\boldsymbol{C}^*_{ij})]=\psi_0+\sum_{s=1}^{S}\psi_s c^*_{ijs}+u_{xi}, \tag{10}$$

where c^*_{ijs} is the value of the s-th (of S) covariates on subject i at time j, and we assume that the random effects u_{xi} are independent and identically distributed $N(0,\sigma^2_{ux})$ variates. One can show that models (9) and (10) imply a compound-symmetric correlation structure for the repeated measures if no predictors are time-dependent. When one or more predictors are time-dependent [certainly the case with X_{ij} in (9)], then the implied correlation between any two repeated measures becomes dependent upon the values of the time-varying predictor(s) at the respective time points. As before, the covariates $\boldsymbol{C}^*$ in (3) may or may not overlap those comprising $\boldsymbol{C}$ in (9).

To set the stage for handling pooled samples to assess exposure (X), first consider the joint likelihood contribution attributable to subject i:

$$\begin{aligned}\Pr\,(\boldsymbol{Y}_i,\boldsymbol{X}_i|\boldsymbol{C}_i,\boldsymbol{C}^*_i)] &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\Pr(Y_i,X_i|u_{yi},u_{xi},\boldsymbol{C}_i,\boldsymbol{C}^*_i)f(u_{yi},u_{xi})\mathrm{d}u_{yi}\mathrm{d}u_{xi}\\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\prod_{j=1}^{n_i}\left[\Pr(Y_{ij}|X_{ij},\boldsymbol{C}_{ij},u_{yi})\times\Pr(X_{ij}|\boldsymbol{C}^*_{ij},u_{xi})\right]f(u_{yi},u_{xi})\mathrm{d}u_{yi}\mathrm{d}u_{xi}.\end{aligned} \tag{11}$$

With normal random effects, such a likelihood is readily specified for maximization using standard statistical software (e.g., the SAS NLMIXED procedure [20]). Note that it is possible to include an additional variance component (σ_{yx}) allowing for correlation between the random effects u_{yi} and u_{xi}. Such residual association between Y and X should seldom be present given that x_{ij} appears as a time-dependent covariate in (2). It may nevertheless be beneficial in practice to assess σ_{yx} via a likelihood ratio test because simulations (not shown) demonstrate that failure to do so can lead to bias in the ML estimate of β.

We assume that pooling of samples is performed within subjects. That is, the n_i specimens that would be assayed to determine the individual repeated binary exposure values $(x_{ij},\ j = 1, \ldots, n_i)$ for a given subject are combined into a single pool for laboratory assessment. As before, we first define the probability (or likelihood contribution) attributable to a negative pool, conditional on the random effects, as follows:

$$\begin{aligned} p_{\boldsymbol{y}i\boldsymbol{0}} &= \Pr\left(\boldsymbol{Y}_i = \boldsymbol{y}_i, \boldsymbol{X}_i = \boldsymbol{0} | u_{yi}, u_{xi}, \boldsymbol{C}_i, \boldsymbol{C}_i^*)\right] \\ &= \prod_{j=1}^{n_i} \left[\Pr(Y_{ij} = y_{ij} | X_{ij} = 0, \boldsymbol{C}_{ij} = \boldsymbol{c}_{ij}, u_{yi}) \times \Pr(X_{ij} = 0 | \boldsymbol{C}_{ij}^* = \boldsymbol{c}_{ij}^*, u_{xi})\right]. \end{aligned} \tag{12}$$

In contrast, a positive pool for subject i makes the contribution $(p_{\boldsymbol{y}i} - p_{\boldsymbol{y}i\boldsymbol{0}})$, where

$$\begin{aligned} p_{\boldsymbol{y}i} &= \Pr\left(\boldsymbol{Y}_i = \boldsymbol{y}_i | u_{yi}, u_{xi}, \boldsymbol{C}_i, \boldsymbol{C}_i^*)\right] \\ &= \sum_{x_{i1}=0}^{1} \sum_{x_{i2}=0}^{1} \cdots \sum_{x_{in_i}=0}^{1} \left\{ \prod_{j=1}^{n_i} \left[\Pr(Y_{ij} = y_{ij} | X_{ij} = x_{ij}, \boldsymbol{C}_{ij} = \boldsymbol{c}_{ij}, u_{yi}) \right.\right. \\ &\qquad \left.\left. \times \Pr(X_{ij} = x_{ij} | \boldsymbol{C}_{ij}^* = \boldsymbol{c}_{ij}^*, u_{xi})\right] \right\}. \end{aligned} \tag{13}$$

The unconditional likelihood contributions corresponding to negative and positive pools are obtained by integrating out the random effects as in (11).

Chen *et al.* [14] proposed maximum likelihood estimation under nonlinear mixed models for pooled binary outcome data via two methods based on a Newton–Raphson and an expectation - maximization (EM) algorithm. The approach given here permits a similar treatment of the case in which the repeated or longitudinally measured binary exposure variable (X) is determined by pooling. After formatting the data such that each record captures the information for a single pool, we have found it convenient to maximize the joint likelihood based on the product of the pool-wise contributions in (12) and (13) via use of the 'general' log-likelihood facility in the SAS NLMIXED procedure [20]. Note that pool size (n_i) equals the number of observations on an individual and is free to vary across subjects. As in the univariate case, 'hybrid' designs with X measured separately on each occasion for certain subjects are also readily accommodated.

3. Practical considerations in the univariate case

3.1. *Applicability in case-control studies*

The joint modeling approach targeting the distribution of (Y, X) given $\boldsymbol{C}$ may appear best suited for application to cross-sectional studies. However, it is well known that, except for distortion of the intercept (α), model (1) can be applied directly under the outcome-dependent sampling typical of case-control studies [16]. In the absence of pooling, it follows that the valid ML estimation of the primary parameters (β, $\boldsymbol{\gamma}$) is also preserved under the joint modeling strategy in (4). The effect of case oversampling will generally be to invalidate the ML estimate of $\boldsymbol{\psi}$, the vector of secondary regression coefficients in model (3).

With pooling to determine exposure status (X), it is useful and convenient to note that the same conclusions hold with regard to the effects of case oversampling. That is, maximization of the likelihood based on (7) and (8) leads to invalid estimates of α and $\boldsymbol{\psi}$, but valid MLEs of the parameters of main interest (β, $\boldsymbol{\gamma}$). This result hinges largely on an appropriate specification of model (3) (refer to Section 3.3), but it is significant in terms of generalizing the approach advocated here beyond the confines of the cross-sectional design. A simulation-based illustration is provided in Section 5 (Table III).

3.2. *Efficient pooling*

In the context of model (1) with Y (rather than X) assessed by pooling, Vansteelandt *et al.* [7] discussed efficiency gains attainable by the informed allocation of subjects into pools. That is, rather than random pooling, they advocate the creation of 'covariate-homogeneous' pools. Similar advice applies to the problem considered here if all subjects are to be allocated to pools for assessing X, except that we recommend the formation of '$(Y,\boldsymbol{C})$-homogeneous' pools. Although the validity of the ML analysis based on (7) and (8) is not contingent on pooling within Y strata (e.g., pooling cases with cases and controls with controls in a case-control study), this strategy will typically be beneficial in the interest of more precise estimation of β. For precision in estimating the remaining coefficients ($\boldsymbol{\gamma}$) in model (1), we recommend sorting subjects within the two strata of Y according to what is viewed as the most scientifically important covariate (C_j), followed by the next most important ($C_{j'}$), and so on. One would then pool together adjacent subjects based on the resulting ordered list, as recommended by prior authors [7].

Importantly, we advocate a slight adjustment to the above strategy in the case of a 'hybrid' design [12], in which a subset of the subjects have X individually measured. In that case, it is generally advantageous to apply sorting by (Y, $\boldsymbol{C}$) not only with a view towards obtaining '(Y,$\boldsymbol{C}$)-homogeneous' pools, but also so as to increase the expected number of negative pools. To clarify, assume that β is expected to be positive in model (1) and that model (3) contains a single covariate C^* whose ψ coefficient is expected to be negative. For a hybrid design in that case, we recommend sorting by Y from the smallest to the largest and then (within the Y strata) sorting by C^* from the largest to the smallest. Pools would be formed from the 'top–down' based on the resulting sorted list so as to tend to increase the number of negative pools obtained, while those nearer the bottom of the list would be assessed individually for X. Simulations that help to illustrate the benefits of informed pooling in both 'all pooled' and 'hybrid' designs are summarized in Section 5 (Tables II and IV, respectively).

3.3. *Effects of misspecifying model (3)*

As mentioned in Section 2, misspecification of model (3) would be of little concern in the absence of pooling assuming that one's main interest lies in the parameters (β, $\boldsymbol{\gamma}$) of model (1). Not surprisingly, the same will not generally be true when applying ML based on (7) and (8) in the case of pooled exposure data. In particular, the judicious choice of covariates ($\boldsymbol{C}^*$) in model (3) becomes important in practice, to ensure valid estimates of (β, $\boldsymbol{\gamma}$) in model (1).

Interestingly, however, there does appear to be a certain robustness to misspecification of model (3). Specifically, we do not expect the MLE of (β, $\boldsymbol{\gamma}$) to be sensitive to the omission of an important covariate (C_j^*) from model (3), provided that C_j^* is not a necessary predictor in model (1). A small simulation study under large sample conditions is provided in Section 5 to illustrate this point (Table V). Nevertheless, we recommend careful selection of the covariates ($\boldsymbol{C}^*$) in model (3) and the usual attentiveness to covariate selection in model (1). Likelihood ratio tests for these purposes are available in the context of the overall likelihood based on (7) and (8).

4. Example

We illustrate the methods for the univariate case in Section 2.1, using data from The Molecular Epidemiology of Colorectal Cancer (MECC) Study. This is a population-based case-control study of all incident cases of colorectal cancer in Northern Israel between March 31, 1998 and March 31, 2004. Details of the study, which collected a rich set of genetic, lifestyle, and environmental variables on over 4000 subjects, can be found elsewhere [21]. For the purposes of model (1), the outcome (Y) is colorectal cancer status and the binary exposure variable (X) is the presence/absence of the C allele at SNP RS16892766 (on chromosomal locus **8q23.3**), which has been found associated with the risk of colorectal cancer via genome-wide association studies making use of Illumina HumanHap BeadChip arrays [21]. Binary control variables ($\boldsymbol{C}$) include other established risk factors for colorectal cancer: statin use (C_1); aspirin use (C_2); average daily consumption of $\geqslant 5$ vegetables (C_3); sports participation (C_4); family history of colorectal cancer (C_5); and ethnicity (Ashkenazi vs other descent; C_6). Age in years (C_7) was treated as a continuous covariate in our analyses.

Here, we utilize data from 1568 cases and 1752 controls between the ages of 26 and 97, with complete data on each of the aforementioned variables. Among these individuals, the overall prevalence of the C allele was approximately 28%. To aid in the illustration, exclusions of a very small number of subjects on the extreme low and high ends of the age range were made to produce sample sizes in both case and control groups that were divisible by eight. Not surprisingly, the presence of the C allele was found to be unassociated with the other model (1) covariates, so we employed an intercept-only version of model (3).

Table I summarizes the complete data results and three additional analyses in which subjects were artificially allocated at random into pools of 2, 4, and 8 within case status (Y) groups. Note the similarity of the point estimate ($\hat{\beta}$) in each case, along with its expected increasing standard error as pool size increases. The loss of precision is nonetheless generally small, especially for pool sizes of 2 or 4. Also of interest is the fact that the MLEs and their standard errors for the $\boldsymbol{\gamma}$ coefficients via the joint

Table I. Results from case-control study associating SNP RS16892766 with colorectal cancer, adjusting for other covariates associated with the disease.*

Predictor variable	Complete data	Pools of 2†	Pools of 4†	Pools of 8†
Presence of C allele (X)	0.312	0.298	0.302	0.383
	(0.079)	(0.087)	(0.108)	(0.179)
Statin use	−0.614	−0.610	−0.617	−0.619
	(0.133)	(0.133)	(0.133)	(0.134)
Aspirin use	−0.432	−0.435	−0.432	−0.433
	(0.095)	(0.096)	(0.096)	(0.096)
Vegetable consumption ($\geqslant$ 5/day)	−0.318	−0.314	−0.319	−0.322
	(0.074)	(0.075)	(0.075)	(0.075)
Sports participation	−0.440	−0.442	−0.439	−0.439
	(0.077)	(0.077)	(0.077)	(0.077)
Family history	0.494	0.490	0.485	0.485
	(0.124)	(0.124)	(0.124)	(0.125)
Age (years)	−0.014	−0.015	−0.015	−0.015
	(0.003)	(0.003)	(0.003)	(0.003)
Ethnicity (Ashkenazi)	0.383	0.388	0.389	0.392
	(0.080)	(0.080)	(0.080)	(0.081)

*Data on 1568 cases and 1752 controls from MECC Study [21, 22]; cells contain ML estimates of regression coefficients and standard errors.

†ML based on Equations (7) and (8); artificial pooling, by case status.

likelihood based on (7) and (8) are almost identical in every case, even when pool sizes increase to 8. This illustrates a striking prospect for resource savings by means of pooling. In particular, it indicates a potential for minimal loss of efficiency with respect to estimating adjusted log ORs corresponding to variables other than X.

All analyses confirm the association of the SNP in question with colorectal cancer in the MECC study, while reiterating the association of other covariates in the model with colorectal cancer.

5. Simulation studies

In this section, we describe simulation studies undertaken to evaluate various aspects of the proposed approaches discussed in Sections 2.1 and 2.2. Except where otherwise noted, in all cases a covariate C in model (1) was generated as a normal variate with standard deviation of 0.5, and C was also utilized as the sole predictor (C^*) in model (3).

5.1. Cross-sectional design with all subjects pooled

Table II summarizes simulations under cross-sectional sampling for ‘all pooled’ designs with various pool sizes and pooling strategies. In each case, the version of model (1) via which data were generated contained the continuous covariate (C) in addition to the binary variable (X). A total of 1000 simulations were run under each of the three scenarios presented, and Table II provides the sample sizes, true parameter values (α, β, γ_1, ψ_0, ψ_1), and pool sizes employed. Under each scenario, three pooling strategies were evaluated for assessing X: (i) random pooling; (ii) pooling after first sorting records by Y; and (iii) pooling after first sorting by Y and by C within Y.

In general, the results show that the ML method in Section 2.1 performs well for pool sizes of 2, 4, and 8. Sorting by Y improves the estimation of β in each case, while sorting by both Y and C preserves validity and improves the estimates of both β and γ_1. Focusing on the ‘Sorted by Y and C’ column, note that modest and miniscule amounts of precision are lost for estimating β and γ_1, respectively, when we analyze 100 pools of size 4 as opposed to 200 pools of size 2. Interestingly, however, utilizing 100 pooled samples with total sample sizes of either 400 or 800 provides similar precision for estimating γ_1, while significant information about β is lost as a result of the larger pool sizes in the latter case. An implication is that the use of larger pools can be beneficial up to a point, but may be subject to diminishing returns. As the footnotes of Table II indicate, another potential disadvantage of larger pool sizes is the higher potential for encountering numerical instability (primarily seen here under random pooling).

Table II. Results from simulation study assessing MLEs under cross-sectional sampling for various pool sizes.*

$N = 400, \quad \alpha = 0, \quad \beta = 2, \quad \gamma_1 = -0.5, \quad \psi_0 = -1, \quad \psi_1 = -0.5$, 200 pools of size 2

	Random pooling		Sorted by Y		Sorted by Y and C	
Parameter	Mean estimate (SD)	Mean estimated SE	Mean estimate (SD)	Mean estimated SE	Mean estimate (SD)	Mean estimated SE
β	2.08 (0.53)	0.49	2.06 (0.36)	0.35	2.05 (0.35)	0.35
γ_1	−0.52 (0.24)	0.23	−0.52 (0.25)	0.24	−0.52 (0.24)	0.23
ψ_0	−0.98 (0.36)	0.35	−0.98 (0.40)	0.39	−0.98 (0.31)	0.30
ψ_1	−0.53 (0.36)	0.35	−0.52 (0.40)	0.39	−0.52 (0.29)	0.29

$N = 400, \alpha = 0, \beta = 2, \gamma_1 = -0.5, \psi_0 = -1, \psi_1 = -0.5$, 100 pools of size 4[†]

	Random pooling		Sorted by Y		Sorted by Y and C	
Parameter	Mean estimate (SD)	Mean estimated SE	Mean estimate (SD)	Mean estimated SE	Mean estimate (SD)	Mean estimated SE
β	1.96 (0.69)	0.81	2.08 (0.42)	0.40	2.07 (0.41)	0.39
γ_1	−0.51 (0.26)	0.26	−0.51 (0.30)	0.29	−0.49 (0.24)	0.24
ψ_0	−1.01 (0.51)	0.49	−1.03 (0.66)	0.61	−0.98 (0.39)	0.38
ψ_1	−0.50 (0.54)	0.51	−0.49 (0.68)	0.63	−0.51 (0.36)	0.35

$N = 800, \quad \alpha = 0, \quad \beta = 2, \quad \gamma_1 = -0.5, \quad \psi_0 = -1, \quad \psi_1 = -0.5$, 100 pools of size 8[‡]

	Random pooling		Sorted by Y		Sorted by Y and C	
Parameter	Mean estimate (SD)	Mean estimated SE	Mean estimate (SD)	Mean estimated SE	Mean estimate (SD)	Mean estimated SE
β	1.95 (1.17)	2.15	2.32 (1.02)	0.72	2.15 (0.62)	0.58
γ_1	−0.53 (0.24)	0.23	−0.53 (0.46)	0.32	−0.49 (0.22)	0.20
ψ_0	−1.03 (0.70)	0.62	−0.92 (0.84)	0.72	−0.89 (0.46)	0.44
ψ_1	−0.50 (0.73)	0.66	−0.52 (0.80)	0.70	−0.54 (0.35)	0.34

SD, standard deviation; SE, standard error.

[*] ML based on Equation (7) and (8); 1000 simulations under each set of conditions.

[†] 37/1000 runs dropped for random pooling from convergence or numerical stability issues.

[‡] 144/1000 runs dropped for random pooling from convergence or numerical stability issues; 45/1000 and 59/1000 runs dropped for 'Sorted by Y' and 'Sorted by Y and C', respectively.

5.2. *Case-control design with all subjects pooled*

In Table III, we summarize simulations designed to mimic a case-control design. First, (Y, X, C) data on a large number of subjects (10,000) were generated according to models (1) and (3), under rare disease conditions. From each of 1000 overall datasets so generated, 200 cases (approximately 50%) and 200 controls (approximately 2%) were randomly selected. Subjects were then allocated into 100 pools of size 4 for assessing X in two ways. First, 50 case pools and 50 control pools were created randomly with respect to C. Second, subjects were sorted by C within case/control status before forming the pools. The true parameter values on which the simulations were based are provided at the top of Table III.

As discussed in Section 3.1, the results indicate maintenance of the validity of the MLEs for (β, γ) via the likelihood based on (7) and (8), while consistency of the (ψ_0, ψ_1) estimators is lost subsequent to case oversampling. Under the conditions in Table III, there was little difference in the precision of the (β, γ) estimates based on pooling by Y only as opposed to sorting by C within Y, although estimates of (ψ_0, ψ_1) were much less variable in the latter case.

5.3. *'Hybrid' design under cross-sectional sampling*

Table IV shows the results of a simulation evaluating the ML approach based on (7) and (8) for a 'hybrid' design in which a total of 600 subjects were simulated in each case, with 300 of them assessed individually for X and the remaining 300 placed into 75 pools of size 4. The true parameter values are given at the top of the table. Note the evidence of improved efficiency for estimating β when pools are formed 'top–down' after first sorting by Y, relative to random pooling. Slight additional precision benefits for estimating β and γ are seen subsequent to sorting by descending values of C within the Y strata, although improvements in the precision of estimates for the secondary parameters (ψ_0 and ψ_1) in model (3) are more pronounced as we move from random to informed pooling scenarios. We refer to Section 3.2 for the explanation of the sorting strategies employed here, based on assumed knowledge of the directionalities of β and γ_1.

5.4. *Effects of omitting a predictor in model (3)*

In Table V, we summarize simulations under large cross-sectional samples to illustrate the impact of omitting an important predictor from model (3) when applying the ML method of Section 2.1. Here, the true model (1) included two independent standard normal covariates (C_1 and C_2) in addition to X, and model (3) also contained two independent standard normal covariates (C_1^* and C_2^*), where $C_1 = C_1^*$. ML estimation based on (7) and (8) was conducted in two ways: (i) utilizing the correct version of

Table III. Results from simulation study assessing MLEs under case-control sampling.*

$N = 400$, $\alpha = -3.5$, $\beta = 1$, $\gamma_1 = -0.5$, $\psi_0 = -2$, $\psi_1 = -0.5$, pools of size 4

	Sorted by Y		Sorted by Y and C	
Parameter	Mean estimate (SD)	Mean estimated SE	Mean estimate (SD)	Mean estimated SE
β	1.04 (0.39)	0.38	1.03 (0.37)	0.37
γ_1	−0.51 (0.22)	0.22	−0.51 (0.21)	0.21
ψ_0	−1.42 (0.62)	0.58	−1.34 (0.36)	0.35
ψ_1	−0.56 (0.70)	0.66	−0.62 (0.37)	0.35

*200 cases, 200 controls per sample; sampling rates of approximately 50% for cases and 2% for controls. ML based on Equations (7) and (8); 1000 simulations.

Table IV. Results from simulation study assessing MLEs under cross-sectional sampling, for a 'hybrid' design.*

$N = 600,\quad \alpha = 0,\quad \beta = 1,\quad \gamma_1 = -0.5,\quad \psi_0 = -1,\quad \psi_1 = -0.5,$ pools of size 4

	Random pooling		Sorted by Y		Sorted by Y and descending C	
Parameter	Mean estimate (SD)	Mean estimated SE	Mean estimate (SD)	Mean estimated SE	Mean estimate (SD)	Mean estimated SE
α	0.001 (0.200)	0.204	0.001 (0.196)	0.199	0.001 (0.195)	0.198
β	1.005 (0.288)	0.288	1.002 (0.237)	0.234	1.001 (0.228)	0.232
γ_1	−0.504 (0.174)	0.177	−0.504 (0.173)	0.176	−0.504 (0.172)	0.175
ψ_0	−1.001 (0.292)	0.278	−0.998 (0.266)	0.257	−1.000 (0.237)	0.228
ψ_1	−0.511 (0.291)	0.276	−0.514 (0.266)	0.254	−0.511 (0.231)	0.220

*Each simulation represents 300 subjects with complete data on X, with the remaining 300 grouped into 75 pools of 4. ML based on Equations (7) and (8); 1000 simulations.

Table V. Results from simulation study assessing effects of omitted predictor in $X|C^*$ model [model (3)].[†]

$N = 50000,\quad \alpha = 0,\quad \beta = 2,\quad \gamma_1 = -0.5,\quad \gamma_2 = 0.25,\quad \psi_0 = -1.5,\quad \psi_1 = -0.5,\quad \psi_2 = 0.5,$ pools of size 4 ($C_2^* = C_2$ is a predictor in both the $Y|X, \boldsymbol{C}$ and $X|\boldsymbol{C}^*$ models)

	Correct $X \quad |\boldsymbol{C}^*$ model	C_2^* omitted from $X \quad |\boldsymbol{C}^*$ model
Parameter	Mean estimate (SD)	Mean estimate (SD)
β	2.00 (0.04)	2.11 (0.04)
γ_1	–0.50 (0.01)	–0.51 (0.01)
γ_2	0.25 (0.01)	0.36 (0.01)

$N = 50000,\quad \alpha = 0,\quad \beta = 2,\quad \gamma_1 = -0.5,\quad \gamma_2 = 0.25,\quad \psi_0 = -1.5,\quad \psi_1 = -0.5,\quad \psi_2 = 0.5,$ pools of size 4 ($C_2^* \neq C_2$, i.e., omitted predictor in $X|\boldsymbol{C}^*$ model is not present in $Y|X, \boldsymbol{C}$ model)

	Correct $X|\boldsymbol{C}^*$ model	C_2^* omitted from $X|\boldsymbol{C}^*$ model
Parameter	Mean estimate (SD)	Mean estimate (SD)
β	1.99 (0.10)	1.98 (0.11)
γ_1	–0.50 (0.01)	–0.50 (0.01)
γ_2	0.25 (0.01)	0.25 (0.02)

[†] ML based on Equations (7) and (8); 25 simulations under each set of conditions.

model (3) and (ii) utilizing a version of model (3) with C_2^* omitted. The top half of Table V shows the results of conducting this experiment when $C_2 = C_2^*$, while the bottom half shows the results when $C_2 \neq C_2^*$ [i.e., the predictor omitted from model (3) is a separate variable not contained in model (1)]. Subjects were allocated to pools of size 4 subsequent to sorting by outcome (Y) status, and the true parameter values are provided in the table. Given the small variability in the MLEs based on the large sample size (50,000), the results are based on 25 simulations in each case.

The top section of Table V reveals that omission of C_2^* produces apparent inconsistency in the MLEs for both β and γ_2 when $C_2^* = C_2$. In contrast, the bottom portion of the table reveals no inconsistencies in the MLE for (β, γ) when the omitted predictor C_2^* in model (3) is not among the necessary covariates in model (1). This illustrates points made in Section 3.3.

5.5. *Longitudinal design with pooling within-subjects*

In Table VI, we present the results of simulations designed to assess the implementation and performance of ML in the setting of a repeated measures study based on the methods in Section 2.2. Data were

Table VI. Results from simulation study assessing MLEs under longitudinal sampling.*

$N = 400,\ \alpha = 0,\ \beta = 0.5,\ \gamma_1 = 1,\ \gamma_2 = -0.5,\ \psi_0 = 0,\ \psi_1 = -2,\ \sigma^2_{uy} = 1,\ \sigma^2_{ux} = 1$, 2 time points				
	Complete data		Pools of size 2†	
Parameter	Mean estimate (SD)	Mean estimated SE	Mean estimate (SD)	Mean estimated SE
β	0.49 (0.29)	0.29	0.50 (0.43)	0.44
γ_1	1.00 (0.11)	0.11	1.00 (0.13)	0.13
γ_2	–0.51 (0.19)	0.20	–0.50 (0.19)	0.19
σ^2_{uy}	1.04 (0.50)	0.49	1.03 (0.49)	0.50
$N = 400,\ \alpha = 0,\ \beta = 0.5,\ \gamma_1 = 1,\ \gamma_2 = -0.5,\ \psi_0 = 0,\ \psi_1 = -2,\ \sigma^2_{uy} = 1,\ \sigma^2_{ux} = 1$, 4 time points				
	Complete data		Pools of size 4†	
Parameter	Mean estimate (SD)	Mean estimated SE	Mean estimate (SD)	Mean estimated SE
β	0.50 (0.24)	0.23	0.50 (0.44)	0.46
γ_1	1.00 (0.08)	0.07	1.00 (0.10)	0.10
γ_2	–0.50 (0.06)	0.06	–0.50 (0.06)	0.06
σ^2_{uy}	1.00 (0.26)	0.26	0.98 (0.25)	0.26

*ML based on Equations (12) and (13); 1000 simulations in each case.

†Convergence in 999/1000 and 960/1000 datasets for pool sizes of 2 and 4, respectively, based on implementation using SAS NLMIXED [20].

simulated under the following versions of models (9) and (10), respectively:

$$\text{logit}[\Pr(Y_{ij} = 1|X_{ij}, C_i, u_{yi})] = \alpha + \beta x_{ij} + \gamma_1 c_i + \gamma_2 t + u_{yi}$$

and

$$\text{logit}[\Pr(X_{ij} = 1|C_i^*)] = \psi_0 + \psi_1 c_i^* + u_{xi},$$

where $C_i = C_i^*$ was a simulated $N(1{,}4)$ baseline (not time-dependent) covariate and $t = j{-}1$ indexes the $j-$th repeated observation. The overall sample size for each simulated dataset was 400, and within-subject pooling for the ascertainment of X was implemented. We conducted the experiment separately for the case of two repeated measures per subject (pool sizes of 2; top half of Table VI), and for the case of four repeated measures per subject (pool sizes of 4; bottom half of Table VI). The true parameter values under which data were generated are given in the table.

The results indicate that ML accounting for the pooling of samples within subjects via Equations (11) and (12) performed well, with few convergence problems based on a convenient implementation facilitated by the SAS NLMIXED procedure ([20]; footnote of Table VI). While the precision of the estimates suffers relative to the complete data analysis as expected, the results under pooling are based on only 50% (for pools of 2) or 25% (for pools of 4) of the total number of required hypothetical laboratory assays, respectively.

Nuisance parameters corresponding to the X given C^* model (3) were estimated reliably with the exception of σ_{ux}^2, for which the MLE exhibited some instability that had little impact on the primary parameter estimates. We note in passing that this issue was readily resolved in further simulations (not summarized here) under similar conditions, in which we employed a hybrid design with X measured directly on each occasion for half of the subjects and via pools for the others. Note also that the simulations summarized in Table VI are based on data and models conforming to the assumption that the random effects (u_{yi} and u_{xi}) are uncorrelated. Other results (not shown) support the feasibility of allowing for correlated errors and time-dependent covariates (C_{ij}).

6. Discussion

We have presented and evaluated ML methods for regression analyses when binary exposure status is assessed via the pooling of biological samples. As discussed by prior authors [5–13], pooling is often appealing as a strategy to reduce the occurrence of assay nondetects and more generally from the standpoint of cost efficiency when assay costs are substantial. In this paper, the methodology proposed for analyzing pooled samples targeting exposure status parallels approaches in the literature for pooled binary outcome assessments in univariate [7] and multivariate [14] settings. The primary distinction is the incorporation of a second regression model for exposure status given covariates [models (3) and (10)], and the need for complete enumeration of the possible individual outcomes when constructing the likelihood for a positive pool [Equations (8) and (13)]. An advantage of the implementations developed here is their computational accessibility by means of general optimization facilities found in standard statistical software [17, 20]. As such, programs for conducting the analyses illustrated in this report are readily available from the authors by request.

The approach taken here provides an appealing and efficient way to incorporate all available information into the analysis, provided that the secondary model for exposure given covariates [(3) or (10)] can be well formulated. Nevertheless, the potential for further research related to pooling scenarios similar to those considered here appears strong. One might seek extensions (perhaps in the context of Section 2.2) that would permit the simultaneous analysis of multiple SNPs assayed via common pooled biological specimens, and/or that would reliably facilitate the incorporation of multiple subject-specific random effects into models (9) and (10) when pooling is to be accommodated. Also in the multivariate case, it may be of interest to consider pooling across (rather than within) subjects, raising new computational challenges because of the need to integrate out multiple random effects for each pool. However, current research in a different context [23] suggests that within-subject pooling may not only be more computationally feasible, but also more efficient statistically.

While the proposed methods are most appealing when accurate (but expensive) lab assays are available, accounting for potential misclassification error in pool-based assessments may also be of interest. Prior studies [7, 14] demonstrate simple adjustments to likelihoods for pooled univariate or multivariate binary response data when the sensitivity and specificity associated with the pool-wise assessment are known, and such an accommodation could be made similarly in the settings considered here. We also note that with a few adjustments to facilitate computations, our approach can be modified to efficiently incorporate more specific information than pool-wise exposure status (e.g., if the assay can accurately determine how many of the pooled subjects are positive). Finally, while Section 3.2 provides practical advice aimed at efficient pooling strategies, further research into more targeted cost-efficiency considerations under variations of 'hybrid' designs [12] that dictate variation in pool sizes based on outcome and covariate information could be of special interest.

Acknowledgements

This research was partially supported by the Long-Range Research Initiative of the American Chemistry Council (ACC) and the Intramural Research Program of the *Eunice Kennedy Shriver* National Institute of Child Health and Human Development (NICHD), National Institutes of Health. Other sources of support include an RC4 grant through the National Institute of Nursing Research (1RC4NR012527-01), an R01 from the National Institute of Environmental Health Sciences (5R01ES012458-07), an RO1 from the National Cancer Institute in support of the MECC Study (1R01CA81488), and a PHS grant (UL 1 RR025008) from the Clinical and Translational Science Award Program, National Institutes of Health, Center for Research Resources. Views expressed are those of the authors and not necessarily of the funding sponsors. The authors gratefully acknowledge the MECC principal investigators Gad Rennert and Stephen B. Gruber and all MECC participants for providing the data resource. We thank each member of the NICHD/ACC-sponsored Methods Development for Biomarkers Working Group, and particularly its organizers (Drs. Enrique Schisterman and Paul Albert).

References

1. Dorfman R. The detection of defective members of a large population. *Annals of Mathematical Statistics* 1943; **14**:436–440.
2. Sobel M, Elashoff R. Group testing with a new goal: estimation. *Biometrika* 1975; **62**:181–193.
3. Boswell MT, Gore SD, Lovison G, Patil GP. Annotated bibliography of composite sampling Part A: 1936-92. *Environmental and Ecological Statistics* 1996; **3**:1–50.
4. Van Belle G, Griffith WC, Edland SD. Contributions to composite sampling. *Environmental and Ecological Statistics* 2001; **8**:171–180.
5. Weinberg CR, Umbach DM. Using pooled exposure assessment to improve efficiency in case-control studies. *Biometrics* 1999; **55**:718–726.
6. Brookmeyer R. Analysis of multistage pooling studies of biological specimens for estimating disease incidence and prevalence. *Biometrics* 1999; **55**:608–612.
7. Vansteelandt S, Goetghebeur E, Verstraeten T. Regression models for disease prevalence with diagnostic tests on pools of serum samples. *Biometrics* 2000; **56**:1126–1133.
8. Liu A, Schisterman EF. Comparison of diagnostic accuracy of biomarkers with pooled assessments. *Biometrical Journal* 2003; **45**:631–644.
9. Mumford SL, Schisterman EF, Vexler A, Liu A. Pooling biospecimens and limits of detection: effects on ROC curve analysis. *Biostatistics* 2006; **7**:585–598.
10. Schisterman EF, Vexler A. To pool or not to pool, from whether to when: applications of pooling to biospecimens subject to a limit of detection. *Pediatric and Perinatal Epidemiology* 2008; **22**:486–496.
11. Vexler A, Schisterman EF, Liu A. Estimation of ROC curves based on stably distributed biomarkers subject to measurement error and pooling mixtures. *Statistics in Medicine* 2008; **27**:280–296.
12. Schisterman EF, Vexler A, Mumford SL, Perkins NJ. Hybrid pooled-unpooled design for cost-efficient measurement of biomarkers. *Statistics in Medicine* 2010; **29**:597–613.
13. Zhang Z, Albert PS. Binary regression analysis with pooled exposure measurements: A regression calibration approach. *Biometrics* 2011; **67**:636–645.
14. Chen P, Tebbs JM, Bilder CR. Group testing regression models with fixed and random effects. *Biometrics* 2009; **65**:1270–1278.
15. Breslow N, Powers W. Are there two logistic regressions for retrospective studies? *Biometrics* 1978; **34**:100–105.
16. Prentice RL, Pyke R. Logistic disease incidence models and case-control studies. *Biometrika* 1979; **66**:403–411.
17. SAS Institute, Inc. *SAS IML 9.1User's Guide*. SAS Institute: Cary, NC, 2004.
18. Diggle PJ, Liang K-Y, Zeger SL. *Analysis of Longitudinal Data*. Oxford University Press: New York, 1994.
19. Davidian M, Giltinan DM. *Nonlinear Models for Repeated Measurement Data*. Chapman & Hall: New York, 1995.
20. SAS Institute, Inc. *SAS STAT 9.1User's Guide*. SAS Institute: Cary, NC, 2004.
21. Poynter JN, Gruber SB, Higgins PD, Almog R, Bonner JD, Rennert HS, Low M, Greenson JK, Rennert G. Statins and the risk of colorectal cancer. *New England Journal of Medicine* 2005; **352**:2184–2192.
22. Houlston RS, Webb E, Broderick P, Pittman AM, Di Bernardo MC, Lubbe S, Chandler I, *et al*. Meta-analysis of genome-wide association data identifies four new susceptibility loci for colorectal cancer. *Nature Genetics* 2008; **40**:1426–1435.
23. Malinovsky Y, Albert PS, Schisterman EF. Pooling designs for outcomes under a Gaussian random effects model. *Biometrics*. DOI: 10.1111/j.1541-0420.2011.01673.x. in press.