Special Issue Paper

Statistics in Medicine

Received 13 January 2011, Accepted 26 July 2011 Published online 23 September 2011 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.4367

Logistic regression analysis of biomarker data subject to pooling and dichotomization

Z. Zhang,[a*†] A. Liu,[a] R. H. Lyles[b] and B. Mukherjee[c]

There is growing interest in pooling specimens across subjects in epidemiologic studies, especially those involving biomarkers. This paper is concerned with regression analysis of epidemiologic data where a binary exposure is subject to pooling and the pooled measurement is dichotomized to indicate either that no subjects in the pool are exposed or that some are exposed, without revealing further information about the exposed subjects in the latter case. The pooling process may be stratified on the disease status (a binary outcome) and possibly other variables but is otherwise assumed random. We propose methods for estimating parameters in a prospective logistic regression model and illustrate these with data from a population-based case-control study of colorectal cancer. Simulation results show that the proposed methods perform reasonably well in realistic settings and that pooling can lead to sizable gains in cost efficiency. We make recommendations with regard to the choice of design for pooled epidemiologic studies. Copyright © 2011 John Wiley & Sons, Ltd.

Keywords: case-control; cohort; cost efficiency; group testing; measurement error; missing covariate

1. Introduction

Biomarkers play a prominent role in epidemiologic research by providing large amounts of valuable information about various exposures. On the other hand, expensive and time-consuming assays are often required to extract information from biomarkers. To make efficient use of limited resources and sometimes to overcome limits of detection, it has become increasingly common in epidemiologic studies to pool specimens across subjects and work with pooled measurements of biomarkers [1–9]. Statistical methods have been developed for regression analysis with pooled exposure measurements in case-control (CC) [1] and cohort studies [9].

All of the aforementioned developments assume that a pooled exposure measurement is an average or, in some cases, a weighted average across subjects. This assumption may be appropriate when the pooling involves equal or known aliquots from different subjects, the mixing process is essentially additive, and the measurement process is sufficiently accurate and precise. However, the assumption can be easily violated if a pooled measurement results from complex biochemical events. For example, allele frequency measurements from pooled DNA samples are subject to errors from several sources: DNA quantification, sample preparation, polymerase chain reaction, and allele frequency determination [10]. Ideally, the statistical analysis should utilize all available information in pooled measurements of allele frequencies, with appropriate modeling techniques to account for the various sources of error. If feasible at all, this would require a lot of modeling assumptions which are difficult to verify in practice. Another approach, which is less than ideal but simpler and presumably more robust, is to dichotomize a pooled measurement. With a single-nucleotide polymorphism (SNP), for instance, it should be easier to ascertain whether a particular allele is present in a pooled DNA sample than to determine its average

[a] *Biostatistics and Bioinformatics Branch, Division of Epidemiology, Statistics and Prevention Research, Eunice Kennedy Shriver National Institute of Child Health and Human Development, Bethesda, MD, USA*
[b] *Department of Biostatistics and Bioinformatics, Rollins School of Public Health, Emory University, Atlanta, GA, USA*
[c] *Department of Biostatistics, School of Public Health, University of Michigan, Ann Arbor, MI, USA*
**Correspondence to: Z. Zhang, BBB/DESPR/NICHD, 6100 Executive Boulevard, Bethesda, MD 20892-7510, USA.*
[†] *E-mail: zhiwei.zhang@nih.gov*

frequency among subjects in the same pool. If the presence of that allele is considered a binary exposure for an individual subject, then a dichotomized measurement of that exposure based on a pooled DNA sample is not an average but the maximum exposure level among individual subjects in the pool.

Interestingly, this dichotomization for a binary exposure subject to pooling is analogous to a group-testing approach to disease screening and prevalence estimation, where subjects are tested in groups and the test result indicates whether the disease or condition is present in the group. The group testing approach has been applied in several areas of medicine including sexually transmitted diseases [11–13], HIV [14–16], hepatitis [17], and drug development [18]. Following decades of research, a variety of statistical methods are now available for screening [11, 18–20], prevalence estimation [21–25], and regression models for disease prediction [26, 27]. All of these methods presume that the variable subject to group testing is the outcome of interest. We are not aware of any previous methodological research on regression problems where group testing is performed on a covariate, such as a SNP in a molecular epidemiologic study.

In this paper, we consider logistic regression analysis where a binary exposure is subject to pooling and the pooled measurement is dichotomized. We describe the study design and formulate the statistical problem in the next section. Then we propose methods for estimating the regression parameters in Sections 3 and 4. It seems necessary to deal with regression coefficients separately for linear terms that do or do not involve the binary exposure subject to pooling. Those involving the exposure subject to pooling can be recovered from a poolwise binary regression model relating the pooled exposure measurement to the outcome and other covariates, as we show in Section 3. For the other regression coefficients, we present two methods in Section 4 whose applicability depends on the pooling mechanism and the nature of the covariates. In Section 5, we illustrate the methods with real data from a CC study of colorectal cancer and evaluate them in simulation experiments mimicking the same study. A discussion in Section 6 concludes the paper.

2. Notation and assumptions

Let Y denote a binary outcome variable such as a disease, X a binary exposure subject to pooling, and Z a collection of confounders and/or effect modifiers. For example, Y may indicate the presence of colorectal cancer, X may represent the presence of a risk allele at a particular SNP, and Z may include environmental variables such as aspirin use, statin use, and vegetable consumption. Suppose that, prospectively, these variables are related through the following logistic regression model:

$$\text{logit P}(Y = 1|X, Z) = \alpha + \beta X + \gamma' Z + \phi'(XZ_a), \tag{1}$$

where Z_a is typically a subvector of Z. The interaction term here is flexible and optional. For a model without interactions, most of our methodology remains applicable if we just ignore the terms involving Z_a or ϕ, with the exception of Section 4.1. We are primarily interested in the association of Y with X while adjusting for Z, which can be characterized by the log-odds ratio $\beta + \phi' Z_a$. We are also interested in the association of Y with Z, which involves γ and ϕ.

We consider both prospective and retrospective (i.e., CC) designs. It is well known that the prospective model (1) remains valid under CC sampling, except for the value of the intercept [28]. We can obtain consistent and efficient estimates of (β, γ, ϕ) by fitting model (1) directly to cross-sectional, cohort, and CC data if all variables are completely observed. The challenge for us is, of course, that X is not measured for each individual subject but rather for pooled specimens only. In this paper, we consider three types of pooling mechanisms:

Random pooling.	Subjects are pooled randomly regardless of their disease status and covariate values.
CC pooling.	Subjects are pooled randomly within each disease status (i.e., cases with cases, controls with controls).
Further stratified (FS) pooling.	The sample is stratified further by Y and Z together, and subjects are pooled randomly within each stratum defined by Y and Z.

We use the subscripts ij to denote the jth subject in the ith pool, $j = 1, \ldots, m_i$, $i = 1, \ldots, n$. It is important to understand the relationship between the subscripted variables (for a particular member of a particular pool) and the unsubscripted ones (for a generic subject chosen randomly from the population).

In a prospective study, random pooling implies that

$$(X_{ij}, Z_{ij}, Y_{ij}) \sim (X, Z, Y), \qquad \text{independently across } (i, j).$$

Obviously, this is not the case under a retrospective design. In the pooling literature, no advantages have been demonstrated for random pooling, as compared with more refined pooling mechanisms, under a retrospective design. We will therefore consider random pooling only for a prospective study. Under CC pooling (regardless of the sampling mechanism), we can make the pooled data representative of the target population by conditioning on the outcome. This can be expressed formally as

$$(X_{ij}, Z_{ij}|Y_{ij} = y_{ij}) \sim (X, Z|Y = y_{ij}), \quad \text{independently across } (i, j). \tag{2}$$

For FS pooling, a connection between pooled and unpooled data can be made by conditioning on (Y_{ij}, Z_{ij}):

$$(X_{ij}|Y_{ij} = y_{ij}, Z_{ij} = z_{ij}) \sim (X|Y = y_{ij}, Z = z_{ij}), \quad \text{independently across } (i, j). \tag{3}$$

Note that the latter two pooling mechanisms can be used in both prospective and retrospective studies, and Equations (2) and (3) hold in both situations because the outcome has already been conditioned upon. In general, it is important to consider the implications of the study design and the pooling mechanism in developing and choosing the appropriate methods of analysis. In the rest of the paper, statements made without specifying the study design or the pooling mechanism will be understood to hold in all possible situations.

We obtain pooled assessments of the X_{ij} as follows. For each pool, we know either that $X_{ij} = 0$ for all j or that $X_{ij} = 1$ for some j; in other words, we observe $V_i = \max\{X_{ij} : j = 1, \ldots, m_i\}$. With the other variables fully observed, we can then summarize the observed data as $(V_i, \mathbf{Y}_i, \mathbf{Z}_i)_{i=1}^n$, where $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{im_i})'$ and $\mathbf{Z}_i = (Z_{i1}, \ldots, Z_{im_i})'$. Valid estimates of (β, γ, ϕ) from such pooled data are not readily available. The naive approach that simply substitutes V_i for X_{ij} is unjustified theoretically and can be shown numerically to yield seriously biased inference. In the next two sections, we propose methods for estimating (β, ϕ) and γ separately.

3. Estimation of (β, ϕ)

The proposed method for estimating (β, ϕ), the exposure-specific coefficients, is based on a poolwise binary regression model for V_i given $(\mathbf{Y}_i, \mathbf{Z}_i)$, which we can derive as follows. We start by writing

$$\mathrm{P}(V_i = 0|\mathbf{Y}_i = \mathbf{y}_i, \mathbf{Z}_i = \mathbf{z}_i) = \prod_{j=1}^{m_i} \mathrm{P}(X_{ij} = 0|Y_{ij} = y_{ij}, Z_{ij} = z_{ij}) = \prod_{j=1}^{m_i} \mathrm{P}(X = 0|Y = y_{ij}, Z = z_{ij}). \tag{4}$$

This is obvious under random pooling. For FS pooling, the above follows directly from expression (3). For CC pooling, it suffices to note that expression (2) implies expression (3), which can be seen as follows. On the event $Y_{ij} = y_{ij}$, the conditional distribution of X_{ij} given Z_{ij} on the left-hand side of (3) is determined by the joint distribution of (X_{ij}, Z_{ij}) on the left-hand side of (2), and the same relationship holds for the unsubscripted variables on the right-hand side. Equality of the joint distributions then implies equality of the conditional distributions.

To understand the far right side of Equation (4), we deduce from model (1) that

$$\mathrm{logit}\mathrm{P}(X = 1|Y, Z) = \beta Y + \phi'(YZ_a) + \log \frac{f(Z|X = 1, Y = 0)}{f(Z|X = 0, Y = 0)} + \text{const.}, \tag{5}$$

where $f(\cdot|\cdot)$ denotes a generic conditional density or mass function. The essence of this relationship has been discussed before without a formal justification [29]. For the reader's convenience, we give a simple proof in the Appendix. The last two terms in (5) are unknown and not straightforward to estimate, so we model them as

$$\log \frac{f(Z|X = 1, Y = 0)}{f(Z|X = 0, Y = 0)} + \text{const.} = \alpha^* + \gamma^{*\prime} Z^*,$$

where Z^* is a (vector-valued) function of Z, which we specify. For example, if Z is a categorical variable, then Z^* is just a collection of dummy variables. If $(Z|X=0,Y=0)$ and $(Z|X=1,Y=0)$ are both normally distributed with the same variance (but possibly different means), then it suffices to set $Z^*=Z$. If the two normal distributions have different variances, then Z^* will also need to involve quadratic terms. In practice, especially when Z has several continuous components, specification of Z^* will require a systematic approach. It is certainly possible to use a general model selection algorithm to compare a number of candidate models (with different specifications of Z^*), either sequentially or simultaneously, in terms of a suitable criterion such as the Akaike Information Criterion or the Bayesian Information Criterion. Another possible criterion, which is specific to this pooling problem, is driven by the objective of estimating (β,ϕ). Using the latter criterion, one would start with a small vector Z^* and add new terms of decreasing importance until the estimates of (β,ϕ) have stabilized. A sensitivity analysis could be performed to compare different starting vectors and different orderings of additional terms.

Once Z^* is specified, we can rewrite (5) as

$$\text{logit P}(X=1|Y,Z)=\alpha^*+\beta Y+\gamma^{*\prime}Z^*+\phi^\prime(YZ_a),$$

and the resulting model for $(V_i|\mathbf{Y}_i,\mathbf{Z}_i)$ is

$$\text{logit P}(V_i=0|\mathbf{Y}_i,\mathbf{Z}_i)=\prod_{j=1}^{m_i}[1+\exp\{\alpha^*+\beta Y_{ij}+\gamma^{*\prime}Z^*_{ij}+\phi^\prime(Y_{ij}Z_{aij})\}]^{-1}, \tag{6}$$

reminiscent of the disease prediction model considered by Vansteelandt et al. [26] in the context of group testing. It is straightforward to estimate $\theta=(\alpha^*,\beta,\gamma^*,\phi)$ by maximizing the likelihood

$$\prod_{i=1}^{n}p(\mathbf{Y}_i,\mathbf{Z}_i;\theta)^{V_i}\{1-p(\mathbf{Y}_i,\mathbf{Z}_i;\theta)\}^{1-V_i},$$

where $p(\mathbf{Y}_i,\mathbf{Z}_i;\theta)=\text{P}(V_i=1|\mathbf{Y}_i,\mathbf{Z}_i)$ is one minus the probability of non-exposure given by (6). The corresponding score equation is

$$\sum_{i=1}^{n}\frac{\{V_i-p(\mathbf{Y}_i,\mathbf{Z}_i;\theta)\}\mathbf{d}(\mathbf{Y}_i,\mathbf{Z}_i;\theta)}{p(\mathbf{Y}_i,\mathbf{Z}_i;\theta)\{1-p(\mathbf{Y}_i,\mathbf{Z}_i;\theta)\}}=\mathbf{0},$$

where

$$\mathbf{d}(\mathbf{Y}_i,\mathbf{Z}_i;\theta)=\sum_{j=1}^{m_i}\frac{\exp\{\alpha^*+\beta Y_{ij}+\gamma^{*\prime}Z^*_{ij}+\phi^\prime(Y_{ij}Z_{aij})\}(1,Y_{ij},Z^{*\prime}_{ij},Y_{ij}Z^\prime_{aij})^\prime}{1+\exp\{\alpha^*+\beta Y_{ij}+\gamma^{*\prime}Z^*_{ij}+\phi^\prime(Y_{ij}Z_{aij})\}}.$$

We have explored a quasi-Newton algorithm as well as an iteratively reweighted least squares algorithm for maximizing the above likelihood. Our numerical experience suggests that the two algorithms yield virtually identical results when they both converge and that the quasi-Newton algorithm is much more stable and less prone to convergence problems. The quasi-Newton algorithm is readily available in R through the 'optim' function (method = 'BFGS'); the only input required is the likelihood and score functions shown above. In addition to the maximum likelihood estimate (MLE), the 'optim' function also returns the observed information.

This method of estimating (β,ϕ) requires specification of the term $\gamma^{*\prime}Z^*$ in model (6), which may appear restrictive at first sight. However, we should note that the method does not require correct specification of the term $\gamma^\prime Z$ in the original model (1); only the terms involving X have an impact on the appearance of model (6). A standard analysis of model (1) based on complete data would require correct specification of the term $\gamma^\prime Z$. Thus, the amount of modeling required for the proposed method is comparable with that for a standard logistic regression analysis of the original, unpooled data.

4. Estimation of γ

We have developed two methods for estimating γ, the regression coefficient for Z in model (1). Their applicability depends on the pooling mechanism and the nature of Z.

4.1. *Maximum likelihood estimate for discrete Z*

Let us start by assuming that Z is discrete, taking integer values between 0 and k, say. We will represent such a Z in model (1) by k indicators (one for each level above 0), and accordingly γ will be a k-vector with elements

$$\gamma_z = \log \frac{\mathrm{P}(Y=1|X=0,Z=z)\mathrm{P}(Y=0|X=0,Z=0)}{\mathrm{P}(Y=0|X=0,Z=z)\mathrm{P}(Y=1|X=0,Z=0)}, \qquad z=1,\ldots,k. \tag{7}$$

We can further express each probability on the right-hand side as

$$\mathrm{P}(Y=y|X=x,Z=z) = \frac{\mathrm{P}(Y=y)\mathrm{P}(Z=z|Y=y)\mathrm{P}(X=x|Y=y,Z=z)}{\sum_{d=0}^{1}\mathrm{P}(Y=d)\mathrm{P}(Z=z|Y=d)\mathrm{P}(X=x|Y=d,Z=z)}.$$

Plugging this into Equation (7), we then have the following characterization:

$$\gamma_z = \log \frac{\mathrm{P}(Z=z|Y=1)\mathrm{P}(Z=0|Y=0)}{\mathrm{P}(Z=0|Y=1)\mathrm{P}(Z=z|Y=0)} + \log \frac{\mathrm{P}(X=0|Y=1,Z=z)\mathrm{P}(X=0|Y=0,Z=0)}{\mathrm{P}(X=0|Y=1,Z=0)\mathrm{P}(X=0|Y=0,Z=z)}. \tag{8}$$

The two terms on the right-hand side pertain to distinct aspects of the underlying distribution: the first term is determined by the conditional distribution of $(Z|Y)$ and the second term by the conditional distribution of $(X|Y,Z)$. This characterization of γ_z has several important implications:

1. It suggests a simple way to estimate γ_z. We can estimate the first term on the right-hand side of (8) by its empirical counterpart

$$\log \frac{N_{1z}N_{00}}{N_{10}N_{0z}}, \tag{9}$$

   where N_{yz} denotes the number of subjects with $Y=y$ and $Z=z$. We can estimate the last term in (8) using the same machinery developed in Section 3. Taking $Z^* = Z_a = Z$, we can rewrite the last term in (8) as

$$\log \frac{\{1+\exp(\alpha^*+\beta)\}\{1+\exp(\alpha^*+\gamma_z^*)\}}{\{1+\exp(\alpha^*+\beta+\gamma_z^*+\phi_z)\}\{1+\exp(\alpha^*)\}}, \tag{10}$$

   where the subscript z denotes the zth element of a vector. We easily obtain an estimate of expression (10) by replacing the unknown parameters with estimates from Section 3. Adding this to (9) yields an estimate of γ_z, which we denote by $\widehat{\gamma}_z$.
2. Equation (8) helps to show that $\widehat{\gamma}_z$ is an MLE. Indeed, because the distributions $(Z|Y)$ and $(X|Y,Z)$ range freely of each other and contribute separately to the likelihood for the observed data, we can obtain their MLEs separately. The estimate given by (9) is clearly an MLE. The estimates given in Section 3 are also MLEs. As a composition of MLEs, $\widehat{\gamma}_z$ must also be an MLE.
3. Equation (8) also helps with variance estimation. The two terms in (8) correspond to uncorrelated score functions, which implies that the variance of $\widehat{\gamma}_z$ is a simple sum of the variance of (9) and the variance of the MLE of (10). Expression (9) has a simple variance estimate: $N_{1z}^{-1}+N_{00}^{-1}+N_{10}^{-1}+N_{0z}^{-1}$, and we can obtain a variance estimate for (10) using the delta method.

To our surprise, extension of this approach to an arbitrary covariate vector Z does not seem straightforward. Intuition suggests that we can generalize the first component of (8) by specifying a model for $(Z|Y)$ or, almost equivalently, for $(Y|Z)$. However, if Z has continuous components, model (1) is generally unsaturated and imposes a subtle constraint on the distributions $(Z|Y)$ and $(X|Y,Z)$. It is not yet clear how to model the latter two distributions in a flexible and convenient way that both complies with model (1) and facilitates estimation of γ. If we can specify these distributions appropriately, it may be possible to deal with the constraint using a profile likelihood approach [30, 31].

4.2. *A restricted Weinberg–Umbach method for case-control pooling*

For a non-discrete Z, estimation of γ is still possible under CC pooling, as we can fit another poolwise binary regression model, for $(\mathbf{Y}_i|\mathbf{X}_i=\mathbf{0},\mathbf{Z}_i)$, using the results of Weinberg and Umbach [1]. Although

$\mathbf{Y}_i$ is a vector, it can only take two values ($\mathbf{0}$ and $\mathbf{1}$) under CC pooling, so this is indeed a binary regression problem. In their original proposal of the CC pooling design, Weinberg and Umbach point out that a logistic regression model for the individual subjects implies a logistic regression model of the same form for pools of subjects with all covariates summed within pools [1]. In the present context, their result can be formulated as

$$\text{logit P}(\mathbf{Y}_i = \mathbf{1}|\mathbf{X}_i, \mathbf{Z}_i) = \log r(m_i) + \alpha^{**} m_i + \beta \sum_{j=1}^{m_i} X_{ij} + \gamma' \sum_{j=1}^{m_i} Z_{ij} + \phi' \sum_{j=1}^{m_i} X_{ij} Z_{aij}, \qquad (11)$$

where α^{**} is a constant that is generally different from α, and $r(m_i)$ is the number of case pools of size m_i divided by the number of control pools of the same size.

Weinberg and Umbach assume that the interaction term is absent and that the X_{ij} in each pool are averaged into a pooled measurement. Together, these two assumptions would allow us to fit model (11) and estimate (β, γ) using a standard logistic regression routine. In this paper, however, we do allow an arbitrary interaction term. More importantly, we work with pooled measurements of the X_{ij} that are obtained as poolwise maxima rather than averages. The information available to us is insufficient for evaluating the sums in model (11) that involve X_{ij}. Nonetheless, estimation of γ is still possible if we restrict attention to pools with $V_i = 0$ (i.e., $\mathbf{X}_i = \mathbf{0}$). This selection process depends only on the covariates in model (11), which therefore remains valid for our selected pools but takes a simpler form:

$$\text{logit P}(\mathbf{Y}_i = \mathbf{1}|\mathbf{X}_i = \mathbf{0}, \mathbf{Z}_i) = \log r(m_i) + \alpha^{**} m_i + \gamma' \sum_{j=1}^{m_i} Z_{ij}. \qquad (12)$$

We can obtain an estimate of γ from a simple logistic regression analysis based on model (12). The restricted Weinberg–Umbach (RWU) method is not available for random pooling and at least not immediately applicable to FS pooling. For CC pooling, it may not be fully efficient because it only uses a portion of the data (i.e., the pools with $V_i = 0$).

5. Numerical results

We now illustrate and evaluate the proposed methods in the setting of a study of colorectal cancer. The Molecular Epidemiology of Colorectal Cancer (MECC) study is a population-based CC study of patients who received a diagnosis of invasive colorectal cancer in northern Israel between March 31, 1998 and March 31, 2004 [32]. Participants were interviewed to obtain demographic information, personal and family history of cancer, medical history, medication use, and health habits. They also completed a dietary questionnaire and had a blood sample collected. Genomic DNA was extracted from blood using the Puregene kit. Genotyping was performed by allele-specific PCR and PCR-restriction fragment length polymorphism. Genotype and other covariate information were available for 3864 subjects (1900 cases and 1964 controls) for our analysis.

We based our comparative analyses of these data and the related simulation experiments on a gene–environment interaction model, where we take X to be the indicator for the presence of a risk allele at a SNP (assuming a dominant model for genetic susceptibility) and Z an environmental variable such as aspirin use (at least one per week in the past 3 years) and sports participation. The SNPs and environmental variables are chosen to cover a range of situations with different probabilities of exposure. Here we consider two choices of X: the C allele at RS16892766 and the G allele at RS548598. The first SNP has been implicated in a genome-wide association study of colorectal cancer [33], whereas the second is chosen to illustrate the rare exposure situation. The proportion of subjects with the allele of interest is approximately 28% for the first SNP and 8% for the second SNP. About 19% of the subjects are aspirin users as defined previously, and 36% of them participate in sports. Both environmental variables have been found to be negatively associated with colorectal cancer, and it will be of interest to see if the association is modified considerably in a subpopulation with a particular genotype. For each choice of (X, Z), the objective is to estimate (β, γ, ϕ) in model (1) with an interaction term.

Table I presents analysis of the MECC data, both in full and in pools, for the first SNP (RS16892766) together with aspirin use. The full-data analysis in the top row reveals a significant positive effect of X, a significant negative effect of Z, and a non-significant interaction between X and Z. The rest of the table is for pooled data, with the leftmost column specifying the pooling mechanism (FS or CC) and the pool

Table I. Analysis of MECC data, both in full and in pools, using a naive approach and the proposed method (PM) with two options for estimating γ (MLE and RWU; see Section 4 for details).

		Point estimate			Standard error		
Pooling	Method	β	γ	ϕ	β	γ	ϕ
None	Full data	0.34	−0.52	−0.15	0.08	0.11	0.20
FS-2	Naive	0.44	−0.50	−0.13	0.08	0.13	0.18
	PM-ML	0.38	−0.53	−0.11	0.09	0.11	0.21
FS-4	Naive	0.50	−0.57	0.00	0.09	0.19	0.21
	PM-ML	0.32	−0.56	0.00	0.11	0.12	0.27
FS-8	Naive	0.72	−0.23	−0.35	0.16	0.33	0.34
	PM-ML	0.29	−0.52	−0.13	0.18	0.14	0.40
CC-2	Naive	0.39	−0.52	−0.06	0.08	0.12	0.18
	PM-ML	0.34	−0.53	−0.08	0.10	0.12	0.31
	PM-RWU	0.34	−0.53	−0.08	0.10	0.13	0.31
CC-4	Naive	0.55	−0.53	−0.01	0.09	0.17	0.20
	PM-ML	0.30	−0.55	0.12	0.14	0.17	0.70
	PM-RWU	0.30	−0.60	0.12	0.14	0.19	0.70
CC-8	Naive	0.62	−0.72	0.17	0.16	0.36	0.37
	PM-ML	0.13	−0.72	0.68	0.26	0.36	1.30
	PM-RWU	0.13	−0.82	0.68	0.26	0.43	1.30

The model includes aspirin use (as Z), the presence of a C allele at RS16892766 (as X), and a possible interaction. The first column specifies the pooling mechanism and the pool size.

size $(2, 4, 8)$. Each pooled sample is analyzed using the proposed method as well as a naive approach that simply substitutes a pooled measurement V_i for each X_{ij} in the pool. The proposed method has two options for estimating γ (MLE and RWU). RWU is only applicable to CC pooling, whereas MLE is applicable to both pooling mechanisms. Under FS-2, the proposed method yields estimates (both point estimates and standard errors) that are fairly close to the full-data analysis. The naive approach yields similar results, too, although its point estimate of β is slightly different. Under FS-4, reasonably close point estimates are obtained for (β, γ) using the proposed method and for γ only under the naive approach. The discrepancies for ϕ are smaller than the corresponding standard errors. Under FS-8, the proposed method again produces point estimates that resemble the full-data analysis and standard errors that are slightly larger as expected, considering the smaller amount of information in the pooled data. The naive approach, on the other hand, produces point estimates that are substantially different from the full-data analysis. The comparison under CC pooling is qualitatively similar, except that no reasonable estimates are produced under CC-8. In most scenarios considered in Table I, the proposed method is consistent with the full-data analysis in terms of the direction and (non-)significance of the covariate effects and the interaction. Larger pool sizes (e.g., 16) have been attempted without producing any informative results due to convergence problems. This is primarily due to difficulties with fitting model (6) when nearly all pools are exposed (i.e., $V_i = 1$), which happens more often with larger pools.

The results based on one or a few pooled samples may be arbitrary and difficult to generalize, so we compare the same methods by using simulated data. Each simulated sample has the exact same numbers of cases and controls as in the original sample, and we generated each simulated sample by sampling cases and controls separately with replacement from the original sample. This mechanism of data generation is consistent with model (1), which is saturated, and the true values of regression parameters are identical to the estimates in a full-data analysis of the original sample (first row of Table I). We analyze each simulated sample using the same set of methods under the same pooling scenarios as in Table I. Simulations for larger pools (FS-16 and CC-16) have not been successful due to convergence problems. For the scenarios shown in Table I, we have encountered no numerical problems under FS pooling and only a minor one under CC pooling (the proposed estimate of γ is not available in one out of 1000 replicate samples under CC-8). The methods are compared in terms of bias, standard deviation (SD), coverage probability (CP) of intended 95% confidence intervals, and relative cost efficiency (RCE), defined as the pool size multiplied by the variance ratio of the full-data analysis to the proposed method. The RCE measures the efficiency of the proposed method relative to the full-data analysis while considering the cost reduction owing to a smaller number of assays for pooled samples. Based on incomplete data, the

proposed method is expected to be less efficient than a full-data analysis, but it requires fewer assays and therefore may be less expensive. The RCE can be a useful measure of cost efficiency if the cost is driven primarily by the assays.

Table II presents the results of a simulation experiment targeted at the analyses in Table I. In the table, it is clear that the naive approach tends to produce biased estimates under both pooling mechanisms. The proposed method is virtually unbiased in small pools and becomes biased for γ and ϕ in large pools (FS-8 and CC-8). In terms of efficiency, the proposed method is nearly as efficient as a full-data analysis under FS-2, slightly less efficient under FS-4, and much less efficient under FS-8. In terms of cost efficiency, the proposed method works best under FS-4. Under CC pooling, the proposed method is generally less efficient than a full-data analysis. The proposed method appears cost efficient (by a small factor) for estimating (β, γ) under CC-2 and CC-4 but not for estimating ϕ. Recall that the two versions of the proposed method differ in estimation of γ under CC pooling. In Table II, they are almost equivalent in small pools in terms of bias and SD, and the difference only shows under CC-8, in which case MLE is less biased and more efficient than RWU. Finally, Table II shows that the proposed method has the desired level of coverage, especially in small pools.

Table III presents pooled and unpooled analyses for the second SNP (RS548598) together with sports participation. The full-data analysis in Table III reveals a significant positive effect of X, a significant negative effect of Z, and a non-significant interaction between X and Z. A notable feature of this SNP is the relatively low occurrence ($\sim$ 8%) of the dominant allele, which implies a smaller probability of poolwise exposure than for the other SNP with the same pool size. This may help to deal with the convergence problem in large pools and allow larger pools to be used. Indeed, Table III shows pooled analyses under FS-16, CC-16, and even FS-32. Other than that, the findings in Table III are somewhat similar to those in Table I. Under both pooling mechanisms, the point estimates from different methods are similar in small pools, and the proposed method continues to produce reasonable point estimates in large pools, at least for (β, γ). Furthermore, the proposed method produces standard errors that seem more plausible, given that a pooled sample contains less information than the original sample.

Table IV presents the results of a simulation experiment targeted at the analyses in Table III. Although Table III includes pooled analyses under FS-32 and CC-16, we have not been able to repeat such analyses

Table II. Empirical comparison of full-data analysis, a naive approach, and the proposed method (PM) with two options for estimating γ (MLE and RWU; see Section 4 for details), in terms of bias, standard deviation (SD), relative cost efficiency (RCE) and coverage probability (CP) of intended 95% confidence intervals, in the setting of the MECC study with X denoting the presence of a C allele at RS16892766 and Z indicating aspirin use.

		Bias			SD			RCE			CP		
Pooling	Method	β	γ	ϕ	β	γ	ϕ	β	γ	ϕ	β	γ	ϕ
None	Full data	0.00	0.00	0.00	0.08	0.11	0.19	1.0	1.0	1.0	0.95	0.95	0.96
FS-2	Naive	0.06	0.04	−0.03	0.11	0.15	0.25				0.77	0.86	0.84
	PM-ML	0.00	0.00	0.00	0.09	0.11	0.21	1.7	1.9	1.7	0.96	0.96	0.95
FS-4	Naive	0.19	0.13	−0.09	0.17	0.32	0.42				0.43	0.67	0.66
	PM-ML	0.00	0.00	0.00	0.11	0.12	0.27	2.2	3.4	2.1	0.95	0.96	0.95
FS-8	Naive	0.51	−0.40	0.47	0.46	3.66	3.69				0.30	0.53	0.53
	PM-ML	0.00	−0.33	0.39	0.18	1.43	1.74	1.8	0.0	0.1	0.96	0.98	0.97
CC-2	Naive	0.04	−0.01	0.07	0.10	0.13	0.18				0.84	0.95	0.94
	PM-ML	0.00	−0.01	0.01	0.10	0.13	0.31	1.4	1.4	0.8	0.96	0.95	0.96
	PM-RWU	0.00	−0.01	0.01	0.10	0.13	0.31	1.4	1.4	0.8	0.96	0.95	0.96
CC-4	Naive	0.15	0.00	0.09	0.17	0.19	0.22				0.52	0.93	0.91
	PM-ML	0.00	0.00	−0.05	0.15	0.19	0.63	1.3	1.3	0.4	0.94	0.93	0.96
	PM-RWU	0.00	0.00	−0.05	0.15	0.19	0.63	1.3	1.2	0.4	0.94	0.94	0.96
CC-8	Naive	0.44	−0.03	0.13	0.42	0.39	0.41				0.34	0.95	0.95
	PM-ML	0.01	−0.04	−0.40	0.27	0.38	2.27	0.8	0.6	0.1	0.95	0.93	0.99
	PM-RWU	0.01	−0.14	−0.40	0.27	1.03	2.27	0.8	0.1	0.1	0.95	0.95	0.99

The true parameter values are based on the full-data analysis in Table I. The first column specifies the pooling mechanism and the pool size. The RCE is calculated by multiplying the pool size with the variance ratio (full-data to PM). Each entry is based on 1000 replicates unless otherwise noted in the text.

Table III. Analysis of MECC data, both in full and in pools, using a naive approach and the proposed method (PM) with two options for estimating γ (MLE and RWU; see Section 4 for details).

		Point estimate			Standard error		
Pooling	Method	β	γ	ϕ	β	γ	ϕ
None	Full data	0.41	−0.41	−0.14	0.16	0.07	0.25
FS-2	Naive	0.38	−0.41	−0.11	0.12	0.08	0.19
	PM-ML	0.37	−0.41	−0.11	0.16	0.07	0.26
FS-4	Naive	0.46	−0.41	−0.08	0.09	0.08	0.15
	PM-ML	0.41	−0.42	−0.08	0.17	0.07	0.27
FS-8	Naive	0.62	−0.20	−0.46	0.08	0.10	0.14
	PM-ML	0.47	−0.39	−0.35	0.18	0.07	0.29
FS-16	Naive	0.48	−0.64	0.21	0.09	0.15	0.17
	PM-ML	0.29	−0.44	0.09	0.22	0.08	0.37
FS-32	Naive	1.59	0.99	−1.50	0.20	0.26	0.27
	PM-ML	0.55	−0.38	−0.45	0.36	0.08	0.55
CC-2	Naive	0.29	−0.16	0.10	0.12	0.05	0.14
	PM-ML	0.22	−0.17	0.21	0.19	0.05	0.26
	PM-RWU	0.22	−0.15	0.21	0.19	0.05	0.26
CC-4	Naive	0.39	−0.43	0.04	0.09	0.08	0.15
	PM-ML	0.29	−0.43	0.16	0.23	0.08	0.59
	PM-RWU	0.29	−0.40	0.16	0.23	0.08	0.59
CC-8	Naive	0.77	−0.44	0.08	0.08	0.10	0.14
	PM-ML	0.38	−0.45	0.67	0.28	0.09	0.97
	PM-RWU	0.38	−0.44	0.67	0.28	0.11	0.97
CC-16	Naive	0.63	−0.41	−0.01	0.10	0.15	0.17
	PM-ML	0.31	−0.42	0.04	0.53	0.15	1.58
	PM-RWU	0.31	−0.55	0.04	0.53	0.20	1.58

The model includes sports participation (as Z), the presence of a G allele at RS548598 (as X), and a possible interaction. The first column specifies the pooling mechanism and the pool size.

Table IV. Empirical comparison of full-data analysis, a naive approach, and the proposed method (PM) with two options for estimating γ (MLE and RWU; see Section 4 for details), in terms of bias, standard deviation (SD), relative cost efficiency (RCE), and coverage probability (CP) of intended 95% confidence intervals, in the setting of the MECC study with X denoting the presence of a G allele at RS548598 and Z indicating sports participation.

		Bias			SD			RCE			CP		
Pooling	Method	β	γ	ϕ	β	γ	ϕ	β	γ	ϕ	β	γ	ϕ
None	Full data	0.00	0.00	−0.01	0.16	0.07	0.26	1.0	1.0	1.0	0.95	0.94	0.94
FS-2	Naive	0.03	0.01	−0.02	0.17	0.08	0.28				0.83	0.93	0.82
	PM-ML	0.01	0.00	−0.02	0.16	0.07	0.27	2.0	2.0	1.9	0.95	0.95	0.94
FS-4	Naive	0.06	0.02	−0.02	0.20	0.12	0.32				0.64	0.82	0.65
	PM-ML	0.01	0.00	−0.01	0.17	0.07	0.28	3.5	4.1	3.4	0.94	0.95	0.94
FS-8	Naive	0.12	0.05	−0.04	0.24	0.21	0.40				0.46	0.60	0.50
	PM-ML	0.00	0.00	−0.01	0.18	0.07	0.30	6.3	7.8	5.9	0.95	0.95	0.95
FS-16	Naive	0.30	0.12	−0.08	0.38	0.53	0.67				0.30	0.40	0.37
	PM-ML	0.01	0.00	0.00	0.22	0.08	0.37	8.8	14.4	7.9	0.96	0.95	0.96
CC-2	Naive	−0.02	0.00	0.06	0.15	0.08	0.20				0.87	0.94	0.93
	PM-ML	0.00	0.00	−0.01	0.20	0.08	0.39	1.4	1.7	0.9	0.94	0.94	0.95
	PM-RWU	0.00	0.00	−0.01	0.20	0.08	0.39	1.4	1.7	0.9	0.94	0.94	0.95
CC-4	Naive	0.00	0.00	0.10	0.16	0.08	0.16				0.73	0.95	0.90
	PM-ML	0.02	0.00	−0.02	0.26	0.08	0.57	1.5	3.2	0.8	0.95	0.95	0.95
	PM-RWU	0.02	0.00	−0.02	0.26	0.08	0.57	1.5	3.0	0.8	0.95	0.95	0.95
CC-8	Naive	0.06	0.00	0.12	0.19	0.10	0.14				0.58	0.95	0.86
	PM-ML	0.02	0.00	−0.03	0.39	0.10	1.01	1.4	4.3	0.5	0.97	0.95	0.98
	PM-RWU	0.02	−0.01	−0.03	0.39	0.11	1.01	1.4	3.7	0.5	0.97	0.95	0.98

The true parameter values are based on the full-data analysis in Table III. The first column specifies the pooling mechanism and the pool size. The RCE is calculated by multiplying the pool size with the variance ratio (full data to PM). Each entry is based on 1000 replicates.

for a large number of simulated samples. For the scenarios shown in Table IV, we have encountered no numerical problems under either pooling mechanism. The findings in Table IV are largely consistent with those in Table II, except that the proposed method performs really well under FS-8 and FS-16, with very little bias and a high RCE for each parameter. This shows that the benefit of pooling can be dramatic for a rare exposure, especially when the pooling is stratified on relevant variables and the pool size is suitably large.

The simulations in Tables II and IV are retrospective in the sense that specified numbers of cases and controls are simulated separately. As suggested by a reviewer, we have also performed prospective simulation experiments where we start with the collection of all subjects (regardless of disease status) with their original covariate values and generate the disease status randomly using the same logistic regression model with parameter estimates from the full-data analysis. This has been done for both gene–environment combinations, and the results are quite similar to those in Tables II and IV and are therefore omitted.

6. Discussion

There is growing interest in pooling specimens in biomarker studies, both as a cost-saving measure and as a way to overcome limits of detection. A fair amount of statistical methodology has been developed for effective use of pooled measurements of biomarkers to address important epidemiologic questions. The available methods for pooled exposure measurements usually assume that a pooled measurement is an average or weighted average of individual exposure levels. This assumption may be questionable in reality, as with pooled DNA samples, and we therefore propose to dichotomize pooled measurements of a binary exposure. The proposed approach avoids complex modeling assumptions about measurement errors and may be preferable when the pooled measurement process is not well understood. On the other hand, it is generally less efficient than methods that incorporate all available information using correctly specified models. Where possible, a sensitivity analysis should be conducted that includes both the proposed approach and non-dichotomized methods based on plausible models.

Numerical results, some of which are not reported here, show that the proposed approach performs reasonably well in several realistic situations. The method for estimating (β, ϕ) may have a convergence problem with large pools; this is more of a design issue than a methodological one (see following paragraphs). There are currently two methods for estimating γ: the MLE method for a discrete Z and the RWU method for CC pooling. When they are both applicable (i.e., CC pooling with a discrete Z), our numerical experience suggests that they perform similarly to each other in small pools, the MLE method being only slightly more efficient. In that case, one may choose to use the RWU method for ease of implementation (although the MLE method is not really difficult to implement). Neither method is applicable under FS pooling with a non-discrete Z, which calls for further research. Another area for further research is how to make joint inference on (β, γ, ϕ) when γ can be estimated. Poolwise bootstrapping seems to be a good possibility.

The findings in this paper demonstrate the cost efficiency of pooled biomarker studies and shed light on the proper design of such studies. We have found that FS pooling generally leads to better efficiency than does CC pooling. This suggests that we should try to identify the most important confounders and effect modifiers in the design stage, stratify on them (in addition to the outcome) in the pooling process, and adjust for them in the subsequent analyses. Since pooling has irreversible consequences, care should be taken to select stratifying variables that are relevant for general purposes and not just in one specific analysis. We have also found that pools should not be too large. The probability of poolwise exposure (i.e., $V_i = 1$) increases with the pool size and the probability of individual exposure (i.e., $X_{ij} = 1$). If the pool size is too large for a given probability of individual exposure, nearly all pools are exposed and the proposed method for estimating (β, ϕ) may fail to converge. This suggests that we should consider the probability of individual exposure in each pooling stratum and make the pool size small enough to ensure a reasonable probability of poolwise exposure. Furthermore, even without the convergence problem, the results in Table II show that the proposed method can perform poorly when the pool size is too large and the number of pools is too small. In practice, we recommend that a conservative pool size be chosen through a range of simulation experiments targeted at the application at hand. A hybrid design where some specimens are pooled and some are not [8] may be a reasonable compromise when pooling seems to help with some analyses but not others.

Appendix A. Proof of Equation (5)

Let us write $h(y,z) = \text{logit}\,\mathrm{P}(X=1|Y=y,Z=z)$. Then $h(1,z)-h(0,z)$ is the log-odds ratio relating X to Y in the subpopulation with $Z=z$. In the same subpopulation, model (1) says that the log-odds ratio relating Y to X is given by $\beta+\phi' z_a$, where z_a is to z what Z_a means to Z. The invariance of odds ratios for binary variables then implies that $h(1,z)-h(0,z)=\beta+\phi' z_a$, and therefore

$$h(y,z) = y(\beta+\phi' z_a) + h(0,z). \tag{13}$$

By definition,

$$h(0,Z) = \text{logit}\,\mathrm{P}(X=1|Y=0,Z) = \log\frac{f(Z|X=1,Y=0)}{f(Z|X=0,Y=0)} + \text{const.}, \tag{14}$$

where the second step follows from Bayes' law. Substituting (14) in (13) completes the proof of (5).

Acknowledgements

This research was supported by the Intramural Research Program of the National Institutes of Health (NIH), *Eunice Kennedy Shriver* National Institute of Child Health and Human Development, and by the Long-Range Research Initiative of the American Chemistry Council. Dr. Lyles' research was partially supported by an R01 grant from the National Institute of Environmental Health Sciences (5R01ES012458-07). The computation was facilitated by the Biowulf cluster computer system made available by the Center for Information Technology at the NIH. We thank Drs. Enrique Schisterman and Paul Albert for helpful discussions and Drs. Stephen B. Gruber and Gad Rennert for sharing the genotype/covariate data from the Molecular Epidemiology of Colorectal Cancer Study supported by NIH grant R01 CA81488.

References

1. Weinberg CR, Umbach DM. Using pooled exposure assessment to improve efficiency in case-control studies. *Biometrics* 1999; **55**:718–726.
2. Faraggi D, Reiser B, Schisterman EF. ROC curve analysis for biomarkers based on pooled assessments. *Statistics in Medicine* 2003; **22**:2515–2527.
3. Liu A, Schisterman EF. Comparison of diagnostic accuracy of biomarkers with pooled assessments. *Biometrical Journal* 2003; **45**:631–644.
4. Mumford SL, Schisterman EF, Vexler A, Liu A. Pooling biospecimens and limits of detection: effects on ROC curve analysis. *Biostatistics* 2006; **7**:585–598.
5. Vexler A, Liu A, Schisterman EF. Efficient design and analysis of biospecimens with measurements subject to detection limit. *Biometrical Journal* 2006; **48**:780–791.
6. 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.
7. 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.
8. 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.
9. Zhang Z, Albert PS. Binary regression analysis with pooled exposure measurements: a regression calibration approach. *Biometrics*. DOI: 10.1111/j.1541-0420.2010.01464.x. in press.
10. Barratt BJ, Payne F, Rance HE, Nutland S, Todd JA, Clayton DG. Identification of the sources of error in allele frequency estimation from pooled DNA indicates an optimal experimental design. *Annals of Human Genetics* 2002; **66**:393–405.
11. Dorfman R. The detection of defective members of large populations. *Annals of Mathematical Statistics* 1943; **14**: 436–440.
12. Kacena K, Quinn S, Hartman S, Quinn T, Gaydos C. Pooling of urine samples for screening for *Neisseria gonorrhoeae* by ligase chain reaction: accuracy and application. *Journal of Clinical Microbiology* 1998; **36**:3624–3628.
13. Kacena K, Quinn S, Howell M, Madico G, Quinn T, Gaydos C. Pooling urine samples for ligase chain reaction screening for genital *Chlamydia trachomatis* infection in asymptomatic women. *Journal of Clinical Microbiology* 1998b; **36**:481–485.
14. Emmanuel JC, Bassett MT, Smith HJ, Jacobs JA. Pooling of sera for human immunodeficiency virus (HIV) testing: an economic method for use in developing countries. *Journal of Clinical Pathology* 1988; **41**:582–585.
15. Cahoon-Young B, Chandler A, Livermore T, Gaudino J. Sensitivity and specificity of pooled vs. individual testing in HIV antibody prevalence study. *Journal of Clinical Microbiology* 1989; **27**:1893–1895.
16. Kline RL, Brother TA, Brookmeyer R, Zeger S. Evaluation of human immunodeficiency virus seroprevalence in population surveys using pooled sera. *Journal of Clinical Microbiology* 1989; **27**:1449–1452.
17. Cardoso M, Koerner K, Kubanek B. Mini-pool screening by nucleic acid testing for hepatitis B virus, hepatitis C virus, and HIV: preliminary results. *Transfusion* 1998; **38**:905–907.

18. Xie M, Tatsuoka K, Sacks J, Young SS. Group testing with blockers and synergism. *Journal of the American Statistical Association* 2001; **96**:92–102.
19. Litvak E, Tu XM, Pagano M. Screening for the presence of a disease by pooling sera samples. *Journal of the American Statistical Association* 1994; **89**:424–434.
20. Gastwirth JL, Johnson WO. Screening with cost-effective quality control: potential applications to HIV and drug testing. *Journal of the American Statistical Association* 1994; **89**:972–981.
21. Sobel M, Elashoff R. Group testing with a new goal: estimation. *Biometrika* 1975; **62**:181–193.
22. Chen CL, Swallow WH. Using group testing to estimate a proportion, and to test the binomial model. *Biometrics* 1990; **46**:1035–1046.
23. Hughes-Oliver JM, Swallow WH. A two-stage adaptive group-testing procedure for estimating small proportions. *Journal of the American Statistical Association* 1994; **89**:982–993.
24. Tu XM, Litvak E, Pagano M. On the informativeness and accuracy of pooled testing in estimating prevalence of a rare disease: application to HIV screening. *Biometrika* 1995; **82**:287–289.
25. Brookmeyer R. Analysis of multistage pooling studies of biological specimens for estimating disease incidence and prevalence. *Biometrics* 1999; **55**:608–612.
26. Vansteelandt S, Goetghebeur E, Verstraeten T. Regression models for disease prevalence with diagnostic tests on pools of serum samples. *Biometrics* 2000; **56**:1126–1133.
27. Chen P, Tebbs JM, Bilder CR. Group testing regression models with fixed and random effects. *Biometrics* 2009; **65**:1270–1278.
28. Prentice RL, Pyke R. Logistic disease incidence models and case-control studies. *Biometrika* 1979; **66**:403–411.
29. Breslow N, Powers W. Are there two logistic regressions for retrospective studies? *Biometrics* 1978; **34**:100–105.
30. Scott AJ, Wild CJ. Maximum likelihood estimation for case-control data. *Biometrika* 1997; **84**:57–71.
31. Chatterjee N, Carroll RJ. Semiparametric maximum likelihood estimation exploiting gene–environment independence in case-control studies. *Biometrika* 2005; **92**:399–418.
32. Poynter JN, Gruber SB, Higgins PDR, 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.
33. Houlston RS, Webb E, Broderick P, Pittman AM, Di B, M C, Lubbe S, Chandler I, Vijayakrishnan J, Sullivan K, Penegar S, Colorectal Cancer Association Study Consortium, Carvajal-Carmona L, Howarth K, Jaeger E, Spain SL, Walther A, Barclay E, Martin L, Gorman M, Domingo E, Teixeira AS, CoRGI Consortium, Kerr D, Cazier JB, Niittymäki I, Tuupanen S, Karhu A, Aaltonen LA, Tomlinson IP, Farrington SM, Tenesa A, Prendergast JG, Barnetson RA, Cetnarskyj R, Porteous ME, Pharoah PD, Koessler T, Hampe J, Buch S, Schafmayer C, Tepel J, Schreiber S, Völzke H, Chang-Claude J, Hoffmeister M, Brenner H, Zanke BW, Montpetit A, Hudson TJ, Gallinger S, International Colorectal Cancer Genetic Association Consortium, Campbell H, Dunlop MG. Meta-analysis of genome-wide association data identifies four new susceptibility loci for colorectal cancer. *Nature Genetics* 2008; **40**:1426–1435.