Efficient Adaptively Weighted Analysis of Secondary Phenotypes in Case-Control Genome-Wide Association Studies

Huilin Li^a Mitchell H. Gail^b

^aDivision of Biostatistics, Department of Population Health, School of Medicine, New York University, New York, N.Y., and ^bDivision of Cancer Epidemiology and Genetics, Biostatistics Branch, National Cancer Institute, NIH, Rockville, Md., USA

Key Words

Adaptively weighted analysis • Case-control study • Genome-wide association study • Maximum likelihood • Secondary phenotype

Abstract

We propose and compare methods of analysis for detecting associations between genotypes of a single nucleotide polymorphism (SNP) and a dichotomous secondary phenotype (X), when the data arise from a case-control study of a primary dichotomous phenotype (D), which is not rare. We considered both a dichotomous genotype (G) as in recessive or dominant models and an additive genetic model based on the number of minor alleles present. To estimate the log odds ratio β_1 relating X to G in the general population, one needs to understand the conditional distribution $[D | X, G]$ in the general population. For the most general model, $[D | X, G]$, one needs external data on $P(D = 1)$ to estimate β_1 . We show that for this ‘full model’, the maximum likelihood (FM) corresponds to a previously proposed weighted logistic regression (WL) approach if G is dichotomous. For the additive model, WL yields results numerically close, but not identical, to those of the maximum likelihood FM. Efficiency can be gained by assuming that $[D | X, G]$ is a logistic model with no interaction between X and G (the ‘reduced model’). However,

the resulting maximum likelihood (RM) can be misleading in the presence of interactions. We therefore propose an adaptively weighted approach (AW) that captures the efficiency of RM but is robust to the occasional SNP that might interact with the secondary phenotype to affect the risk of the primary disease. We study the robustness of FM, WL, RM and AW to misspecification of $P(D = 1)$. In principle, one should be able to estimate β_1 without external information on $P(D = 1)$ under the reduced model. However, our simulations show that the resulting inference is unreliable. Therefore, in practice one needs to introduce external information on $P(D = 1)$, even in the absence of interactions between X and G .

Copyright © 2012 S. Karger AG, Basel

Introduction

Genome-wide association studies (GWAS) usually measure hundreds of thousands of single nucleotide polymorphisms (SNPs) on thousands of subjects with a primary disease (cases) and without the primary disease (controls). Often, data on other characteristics (‘secondary phenotypes’) of cases and controls are available, and researchers wish to study the association between a secondary phenotype and SNPs by taking advantage of the

valuable data already collected for the primary disease. However, the case-control study sample is not representative of the entire population, for which the associations between the secondary phenotype and SNPs are desired. Ignoring the ascertainment in the case-control study design in the secondary analysis can induce biased estimates of association and false-positive significance tests, as described in Jiang et al. [1], Lin and Zeng [2], Richardson et al. [3], Monsees et al. [4] and Wang and Shete [5].

Li et al. [6] studied methods of testing for associations between a SNP from a case-control GWAS and a secondary phenotype in the special case that the primary disease in the GWAS was rare. They assumed that the secondary phenotype ($X = 0$ or 1) and SNP genotype ($G = 0$ or 1) were binary (as for dominant or recessive genetic models) and the primary disease status indicator D was 0 or 1 , according as the primary disease was absent or present. The aim was to estimate the log odds ratio relating X and G in the general population based on

$$\text{logit}(P(X = 1 | G)) = \beta_0 + \beta_1 G. \quad (1)$$

To take account of the case-control sampling, one must consider the model

$$P(D = 1 | G, X) = \frac{\exp(\delta_0 + \delta_1 G + \delta_2 X + \delta_{12} GX)}{1 + \exp(\delta_0 + \delta_1 G + \delta_2 X + \delta_{12} GX)}. \quad (2)$$

For a rare primary disease, the denominator of equation (2) disappears from the likelihoods, simplifying the analysis. Li et al. [6] showed that under the general model (2) and for a rare primary disease, the maximum likelihood estimation of β_1 is equivalent to analyzing data from the controls alone. If one assumes, as in Lin and Zeng [2], that $\delta_{12} = 0$ in equation (2), then a much more efficient estimate of β_1 is obtained, whose variance is only very slightly smaller than obtained by taking a weighted average of estimates from cases and controls, which we call the efficient weighted estimate (see [6]). However, these estimates were biased and led to the above nominal hypothesis test sizes if in fact $\delta_{12} \neq 0$. Therefore, Li et al. [6] proposed an adaptively weighted estimator that put more weight on the control-only estimate when the data suggested $\delta_{12} \neq 0$ and put more weight on the efficient weighted estimate when there was little evidence that $\delta_{12} \neq 0$.

We now consider the case when the primary disease condition is not rare, as might arise in GWAS of visual acuity or elevated blood pressure. Such data might arise, for example, from a hospital-based case-control study. In a hypertension clinic, a sample of cases might be selected for genome-wide scanning and compared to a sample of

patients without hypertension. Such a study would not yield information on the probability of hypertension in the general population. A cohort study might yield a large number of subjects with diminished visual acuity. Genome scans might be performed on some members of the cohort with diminished visual acuity and some members without diminished visual acuity. In this context, the probability of diminished visual acuity, $P(D = 1)$, could be estimated from the cohort information. We compare the adaptively weighted model to other methods for a disease that is not rare, and we treat both dichotomous models for G and additive models in which G represents the number of minor alleles.

For a disease that is not rare, estimates of the other parameters in models (1) and (2) apart from β_1 can be very unstable and lead to invalid inferences for secondary phenotypes unless the disease prevalence is known [2]. Thus, one needs to assume that $P(D = 1)$ is known, in which case β_1 can be estimated by maximum likelihood under equations (1) and (2) to yield $\hat{\beta}_{1FM}$. Here, FM stands for full model. Under this assumption, or assuming that the sampling fractions for cases and controls are known, an apparently simpler estimate can be obtained by reweighting the log-likelihood corresponding to equation (1) to obtain a weighted estimate $\hat{\beta}_{1WL}$ [4]. This estimate had been obtained earlier as a consequence of weighted logistic regression [3]. We prove in this paper that $\hat{\beta}_{1WL}$ is in fact the maximum likelihood estimate (MLE) $\hat{\beta}_{1FM}$ under the dichotomous genetic model. In simulations, we find that $\hat{\beta}_{1WL}$ is numerically very near but not equal to $\hat{\beta}_{1FM}$ for the additive genetic model.

Efficiency can be improved over $\hat{\beta}_{1FM}$ if one is willing to assume $\delta_{12} = 0$ and known $P(D = 1)$. In fact, Lin and Zeng [2] studied this case. We denote the corresponding MLE $\hat{\beta}_{1RM}$, which is more efficient than $\hat{\beta}_{1FM}$. Here, RM stands for reduced model. However, $\hat{\beta}_{1RM}$ can be misleading if $\delta_{12} \neq 0$. Therefore, we developed an adaptive estimate, $\hat{\beta}_{1AW}$, that puts more weight on $\hat{\beta}_{1FM}$ when there is evidence that $\delta_{12} \neq 0$ and more weight on $\hat{\beta}_{1RM}$ when there is less evidence that $\delta_{12} \neq 0$.

In principle, one should be able to estimate β_1 when $\delta_{12} = 0$ by MLE without knowing $P(D = 1)$, as discussed in Lin and Zeng [2]. The corresponding estimate is denoted as $\hat{\beta}_{1RMU}$ in this paper. A potential advantage of $\hat{\beta}_{1RMU}$ is that it does not require specification of $P(D = 1)$, whereas $\hat{\beta}_{1WL}$, $\hat{\beta}_{1FM}$, $\hat{\beta}_{1RM}$ and $\hat{\beta}_{1AW}$ do. Our numerical studies show that for $\delta_{12} = 0$ $\hat{\beta}_{1RMU}$ yields unbiased estimates of β_1 , but estimates of the variance of $\hat{\beta}_{1RMU}$ from the information matrix can be too small, which leads to hypothesis tests with sizes above nominal levels. More-

over, when $\delta_{12} \neq 0$, $\hat{\beta}_{1RMU}$ can be seriously misleading, just as can $\hat{\beta}_{1RM}$. Therefore, methods based on external knowledge of $P(D = 1)$ are needed in practice, and we study the robustness of the estimators $\hat{\beta}_{1WL}$, $\hat{\beta}_{1FM}$, $\hat{\beta}_{1RM}$ and $\hat{\beta}_{1AW}$ to misspecification of $P(D = 1)$. We compare these estimators for studying associations with a pre-selected candidate SNP and for discovering a SNP associated with X among all the SNPs studied in the GWAS data.

In the next section, we describe the methods in more details for a common primary disease. Then, we present results of analyses and numerical studies and, finally, we discuss these results. Most technical details are presented in the Appendices.

Methods

We first consider the important scenario of an unmatched case-control study with a dichotomous genotype G and a secondary phenotype X , but where the primary disease is not rare. We extend all the methods to the additive genetic model at the end of this section. The data for the dichotomous G can be represented as a 2-by-4 array (table 1). Let $r_0 = (r_{000}, r_{001}, r_{010}, r_{011})$ and $r_1 = (r_{100}, r_{101}, r_{110}, r_{111})$ denote the case and control cell frequency vectors, respectively, and let n_0 and n_1 be the numbers of controls and cases selected, respectively. In the table, r_{dgs} represents the cell counts for $D = d$, $G = g$ and $X = s$.

Weighted Logistic Regression Method

Jiang et al. [1] and Richardson et al. [3] proposed a weighted logistic regression method that can be used to estimate the association between genotype and secondary phenotype. Monsees et al. [4] studied this method through simulations in GWAS that assumed $\delta_{12} = 0$. Defining weights that are inversely proportional to the sampling fractions $w_i = P(D = i)/n_i$ for cases ($i = 1$) and controls ($i = 0$), one can construct a weighted pseudo-log-likelihood from the data in table 1 as

$$\begin{aligned} l_w &= \sum_{i=0}^1 w_i \sum_{j=1}^{n_i} \log(P(X_{ij}|G_{ij})) \\ &= \sum_{x=0}^1 \sum_{g=0}^1 (w_0 r_{0gx} + w_1 r_{1gx}) \frac{\exp[(\beta_0 + \beta_1 g)x]}{1 + \exp(\beta_0 + \beta_1 g)} \\ &= w_0 \left[r_{000} \log\left(\frac{1}{1 + \exp(\beta_0)}\right) + r_{010} \log\left(\frac{1}{1 + \exp(\beta_0 + \beta_1)}\right) \right. \\ &\quad \left. + r_{001} \log\left(\frac{\exp(\beta_0)}{1 + \exp(\beta_0)}\right) + r_{011} \log\left(\frac{\exp(\beta_0 + \beta_1)}{1 + \exp(\beta_0 + \beta_1)}\right) \right] \\ &\quad + w_1 \left[r_{100} \log\left(\frac{1}{1 + \exp(\beta_0)}\right) + r_{110} \log\left(\frac{1}{1 + \exp(\beta_0 + \beta_1)}\right) \right. \\ &\quad \left. + r_{101} \log\left(\frac{\exp(\beta_0)}{1 + \exp(\beta_0)}\right) + r_{111} \log\left(\frac{\exp(\beta_0 + \beta_1)}{1 + \exp(\beta_0 + \beta_1)}\right) \right]. \end{aligned} \quad (3)$$

Table 1. Data for an unmatched case-control study with dichotomous genotype and phenotype

	$G = 0$		$G = 1$		Total
	$X = 0$	$X = 1$	$X = 0$	$X = 1$	
$D = 0$	r_{000}	r_{001}	r_{010}	r_{011}	n_0
$D = 1$	r_{100}	r_{101}	r_{110}	r_{111}	n_1

Maximizing equation (3), one obtains the estimate of β_1 :

$$\hat{\beta}_{1WL} = \log \left(\frac{w_0 r_{011} + w_1 r_{111}}{w_0 r_{010} + w_1 r_{110}} \frac{w_0 r_{000} + w_1 r_{100}}{w_0 r_{001} + w_1 r_{101}} \right). \quad (4)$$

The justification for equation (4) is that the corresponding score equations obtained by differentiation of equation (3) with respect to β_0 and β_1 have expectation zero in the entire population, even though the data were obtained from a stratified (on D) random sample from the population. Thus, equation (4) is unbiased for β_1 . This approach requires no modeling of the primary disease probability as in equation (2) and is therefore robust to a possible non-zero interaction δ_{12} between X and G (see Results section). We address how efficient $\hat{\beta}_{1WL}$ is compared to maximum likelihood when the weights are known, which we assume hereafter. Because we usually need to estimate $P(D = 1)$ from external data, we also study the sensitivity of various estimates to misspecification of $P(D = 1)$.

Maximum Likelihood

Jiang et al. [1] discussed efficient semi-parametric likelihood methods for secondary phenotype analysis in case-control studies. Lin and Zeng [2] investigated this method for secondary phenotypes in GWAS. Their method is based on the retrospective likelihood

$$\begin{aligned} L &= \prod_{i=0}^1 \prod_{j=1}^{n_i} P(X_{ij}, G_{ij} | D_{ij} = i) \\ &= \prod_{i=0}^1 \prod_{j=1}^{n_i} P(D_{ij} = i | X_{ij}, G_{ij}) P(X_{ij} | G_{ij}) P(G_{ij}) / P(D_{ij} = i), \end{aligned} \quad (5)$$

where $i = 0, 1$ indexes the primary disease status and $j = 1, \dots, n_i$ indexes the subjects with $D = i$. For binary genotypes and secondary phenotypes as in table 1, equation (5) can be rewritten as

$$\prod_{d=0}^1 \prod_{g=0}^1 \prod_{x=0}^1 \left(\frac{P(D = d | X = x, G = g) P(X = x | G = g) P(G = g)}{P(D = d)} \right)^{r_{dgs}}. \quad (6)$$

In most of this paper, we assume that the disease prevalence $P(D = 1)$ is known, and we let $PG_0 = P(G = 0)$ and $PG_1 = P(G = 1) = 1 - PG_0$. Maximizing equation (6) is equivalent to maximizing

$$\prod_{d=0}^1 \prod_{g=0}^1 \prod_{x=0}^1 \left(P(D = d | X = x, G = g) \frac{\exp(x(\beta_0 + \beta_1 g))}{1 + \exp(\beta_0 + \beta_1 g)} PG_g \right)^{r_{dgs}} \quad (7)$$

subject to the constraint

$$\sum_{x=0}^1 \sum_{g=0}^1 P(D=1|X=x, G=g) \frac{\exp(x(\beta_0 + \beta_1 g))}{1 + \exp(\beta_0 + \beta_1 g)} PG_g$$

$$= P(D=1).$$

Depending on which disease model is used, one can obtain two different MLEs of β_1 from equation (7). If model (2) is used, there will be seven unknown parameters $\theta_{FM} = \{\beta_0, \beta_1, \delta_0, \delta_1, \delta_2, \delta_{12}, PG_0\}$ in the likelihood (7). Their MLEs are denoted as $\hat{\theta}_{FM}$. If $\delta_{12} = 0$ is assumed in model (2), there will be six unknown parameters $\theta_{RM} = \{\beta_0, \beta_1, \delta_0, \delta_1, \delta_2, PG_0\}$ in the likelihood (7), and we denote the corresponding MLEs by $\hat{\theta}_{RM}$. One might anticipate that $\hat{\beta}_{1RM}$ is more efficient than $\hat{\beta}_{1FM}$, and numerical studies discussed in the Results section confirm this hypothesis. Lin and Zeng [2] also estimated β_1 from equation (6) without assuming that $P(D=1)$ is known by setting $\delta_{12} = 0$. We study the properties of their estimator and denote it by $\hat{\beta}_{1RMU}$.

Adaptively Weighted Estimate

To capture the efficiency of $\hat{\beta}_{1RM}$ while avoiding the bias in this estimate that results when $\delta_{12} \neq 0$, we propose an estimator, as in Li et al. [6], that adaptively combines $\hat{\beta}_{1FM}$ and $\hat{\beta}_{1RM}$ as

$$\hat{\beta}_{1AW} = \frac{\hat{\delta}_{12FM}^2}{\hat{\delta}_{12FM}^2 + \hat{\sigma}_{FM}^2} \hat{\beta}_{1FM} + \frac{\hat{\sigma}_{FM}^2}{\hat{\delta}_{12FM}^2 + \hat{\sigma}_{FM}^2} \hat{\beta}_{1RM}. \quad (8)$$

In this equation, $\hat{\delta}_{12FM}^2$ is the MLE of interaction in the full model, and $\hat{\sigma}_{FM}^2$ is the estimated variance of $\hat{\beta}_{1FM}$. This new method strikes a balance between bias and efficiency. It puts more weight on the efficient $\hat{\beta}_{1RM}$ when $\hat{\delta}_{12FM}^2$ is small compared to $\hat{\sigma}_{FM}^2$ and more weight on the less efficient $\hat{\beta}_{1FM}$ when $\hat{\delta}_{12FM}^2$ is large compared to $\hat{\sigma}_{FM}^2$, which indicates that there is an interaction between the effects of secondary phenotypes and genotypes on the risk of the primary disease.

Additive Genetic Model

When an additive genetic model is assumed for the genotype G , we use $G = \{0, 1, 2\}$ to denote the number of minor alleles in the SNP genotype. The data can be represented in a 2-by-6 array in which controls with frequency vector $r_0 = (r_{000}, r_{001}, r_{010}, r_{011}, r_{020}, r_{021})$ and cases with frequency vector $r_1 = (r_{100}, r_{101}, r_{110}, r_{111}, r_{120}, r_{121})$ are in separate rows, similar to table 1. We assume Hardy-Weinberg equilibrium and express $PG_0 = P(G=0) = (1-p)^2$, $PG_1 = P(G=1) = 2p(1-p)$ and $PG_2 = P(G=2) = p^2$, where p is the unknown minor allele frequency. From these data, one can obtain the pseudo-log-likelihood for the weighted logistic regression method as in equation (3) and the retrospective likelihood for the maximum likelihood methods as in equations (5), (6) and (7) where the disease prevalence is assumed known. The adaptively weighted estimate $\hat{\beta}_{1AW}$ is computed from the MLEs $\hat{\beta}_{1FM}$, $\hat{\delta}_{12FM}^2$ and $\hat{\beta}_{1RM}$ from equation (8).

Under the additive genetic model, $\hat{\beta}_{1WL}$ cannot be written explicitly, and iterative numerical methods are needed. We used SAS PROC SURVEYLOGISTIC, which also yields a correct variance estimate. Under model (2), seven unknown parameters $\theta_{FM} = \{\beta_0, \beta_1, \delta_0, \delta_1, \delta_2, \delta_{12}, p\}$ appear in the likelihood (7), and the corresponding MLE is $\hat{\theta}_{FM}$. If $\delta_{12} = 0$ in model (2), there are six unknown parameters $\theta_{RM} = \{\beta_0, \beta_1, \delta_0, \delta_1, \delta_2, p\}$, and we denote the corresponding MLE by $\hat{\theta}_{RM}$.

Results

Analytic Findings

For dichotomous G , the MLE $\hat{\beta}_{1FM}$ can be expressed in closed form as a function of r_{ijk} and $P(D=1)$ (see Appendices). This proves $\hat{\beta}_{1FM} = \hat{\beta}_{1WL}$. Thus, the weighted logistic estimate is fully efficient under the full disease model (2). However, numerical calculations based on the Fisher information and simulation studies described later show that $\hat{\beta}_{1FM}$ is considerably less efficient than $\hat{\beta}_{1RM}$, which is based on the additional assumption $\delta_{12} = 0$. If $P(D=1) \approx 0$, the estimate $\hat{\beta}_{1WL} = \hat{\beta}_{1FM}$ in equation (4) reduces to the log odds ratio in control subjects only. Li et al. [6] had previously shown that for a rare primary disease, $\hat{\beta}_{1FM}$ is the log odds ratio in controls only, but the new results show that $\hat{\beta}_{1FM} = \hat{\beta}_{1WL}$ whether the disease is rare or not. For the additive genetic model, simulations indicate that $\hat{\beta}_{1WL}$ is numerically close, but not identical, to $\hat{\beta}_{1FM}$. Therefore, in what follows we usually present data for $\hat{\beta}_{1FM}$ and not for $\hat{\beta}_{1WL}$.

The Appendices outline algorithms to estimate $\hat{\beta}_{1RM}$ for the case where $P(D=1)$ is known and $\hat{\beta}_{1RMU}$ for the case where $P(D=1)$ is not known. Following re-parameterization, these estimates are closed form functions of certain cell probabilities. However, iterative methods are needed to estimate those cell probabilities. A variance estimate for $\hat{\beta}_{1AW}$ is given in the Appendices that takes the variability of $\hat{\beta}_{1FM}$, $\hat{\beta}_{1RM}$, and $\hat{\delta}_{12FM}^2$ into account. We refer to procedures such as hypothesis tests or confidence interval estimations that are associated with $\hat{\beta}_{1WL}$, $\hat{\beta}_{1FM}$, $\hat{\beta}_{1RM}$, $\hat{\beta}_{1AW}$ and $\hat{\beta}_{1RMU}$ as WL, FM, RM, AW and RMU procedures, respectively.

Simulations to Compare Estimates and Tests for a Pre-Selected SNP

First, we consider the case of a pre-selected SNP, as might arise in a candidate gene study. We used Monte Carlo simulation to evaluate the performance of different estimators and procedures for a pre-selected SNP under the dichotomous genetic model. The simulation results for the additive genetic model are quite similar to those for the dichotomous genetic model and are briefly summarized at the end of this section. For dichotomous G , we fixed the probabilities in the general population of carrying one or two alleles of interest ($G=1$) as $P(G=1) = 0.3$. We set $\beta_0 = 0$ and let $\beta_1 = 0$ and 0.25 under the null and alternative hypotheses, respectively. For the disease model, we set $\delta_1 = 0$ and $\delta_2 = \log(1.5)$, varied the value δ_{12} from -1.5 to 1.5 and choose the value of δ_0 to yield a disease prevalence $P(D=1)$ of 0.10 or 0.30. For each set of simu-

lation parameters, the conditional distributions of X and G given D were determined, and we generated 10,000 datasets with 1,000 cases and 1,000 controls from two independent quadrinomial distributions, $[X, G | D]$, corresponding to the case ($D = 1$) and control ($D = 0$) populations. We obtained estimates $\hat{\beta}_{1FM}$, $\hat{\beta}_{1RM}$, $\hat{\beta}_{1AW}$ and $\hat{\beta}_{1RMU}$, together with their estimated variances from the corresponding information matrices for $\hat{\beta}_{1FM}$, $\hat{\beta}_{1RM}$ and $\hat{\beta}_{1RMU}$. The variance of $\hat{\beta}_{1AW}$ was calculated as in the Appendices from Taylor series expansion. Wald statistics, calculated as the squared estimate divided by its estimated variance, led to rejection of the null hypothesis for values exceeding 3.84, the 95th percentile of the χ^2 distribution with 1 d.f. In order to study the robustness of our procedures to misspecification of $P(D = 1)$, we used values 0.11 and 0.12 instead of the correct $P(D = 1) = 0.10$, and values 0.32 and 0.34, instead of the correct $P(D = 1) = 0.30$.

Figure 1 gives the sizes of the various procedures for testing $H_0: \beta_1 = 0$ for true primary disease probability 10% (fig. 1a–c) and 30% (fig. 1d–f) for the dichotomous genetic model. Correct specification of the primary disease probability corresponds to figure 1a, d, and the effects of misspecification are shown in the middle and right columns. Consider first the case where $P(D = 1)$ is correctly specified. Then the FM and AW tests have nominal size across the range of values of δ_{12} both for $P(D = 1)$ values of 10 and 30%. As expected, the RM test has nominal size when $\delta_{12} = 0$, but not otherwise. The RMU procedure, which does not require specifying $P(D = 1)$, has above nominal size not only when $\delta_{12} \neq 0$, as expected, but also when $\delta_{12} = 0$. This is because estimates of the standard error of variance of $\hat{\beta}_{1RMU}$ are skewed, and many are very small (fig. 2b), leading to many rejections even for small values of $\hat{\beta}_{1RMU}$, as shown by the red dots in figure 2d. In contrast, at $\delta_{12} = 0$, the size of the RM procedure is nominal (fig. 1a, d) and the corresponding estimates of the standard errors of estimates of $\hat{\beta}_{1RM}$ are not skewed (fig. 2a, c). These findings are confirmed by table 2, where the size of the RMU procedure is seen to be near 0.10 at $\delta_{12} = 0$. We conclude that only the FM and AW procedures should be used in general, but that the RM procedure has proper size at $\delta_{12} = 0$. The RMU procedure should not be used because it does not control size, even for a pre-selected SNP. Therefore, we do not present results for the RMU method in the following simulations. Separate simulations show that, even with only 200 cases and 200 controls, FM and AW tests have nominal size (unreported data). Other simulations yielded similar results with $\delta_1 = \log(1.5)$ and $P(X = 1) = 0.5$ (unreported data). The previous conclusions are robust to misspecification of $P(D = 1)$

when the true probability is 10% (fig. 1b, c). When $P(D = 1) = 30\%$, the FM and AW procedures have near nominal size with misspecified values of $P(D = 1)$ for values of δ_{12} in the range from -0.5 to 0.5 . Outside this range, overestimating $P(D = 1)$ results in some elevation in size above nominal levels (fig. 1e, f).

Figure 3 presents the power for the FM and AW procedures. Results are also given for the RM procedure, although the power should be ignored for RM except when $\delta_{12} = 0$, because the size is otherwise not controlled for the RM procedure. At $\delta_{12} = 0$, the RM procedure has higher power than the AW procedure, which has higher power than the FM procedure, both for $P(D = 1) = 10\%$ and $P(D = 1) = 30\%$ (fig. 3a, d). These relationships are shown clearly in table 2. This is not surprising since RM and, to a lesser extent, AW take advantage of the assumption that $\delta_{12} = 0$. For $P(D = 1) = 10\%$, the power of AW exceeds that of FM for $\delta_{12} > -0.2$, but is less than that of FM for $\delta_{12} < -0.2$ (fig. 3a). Note that AW is more efficient than FM in the region $|\delta_{12}| < 0.2$ that surrounds $\delta_{12} = 0$. This property of AW leads to higher power for AW than FM in simulations for GWAS in the next section.

For $P(D = 1) = 30\%$, a similar pattern is seen, but the absolute difference in power between the FM and AW procedures is small (fig. 3d). For $P(D = 1) = 10\%$, the powers of the FM and AW procedures are not greatly changed by misspecification of $P(D = 1)$ (fig. 3b, c). In contrast, for $P(D = 1) = 30\%$, overestimation of $P(D = 1)$ changes the power appreciably for values of δ_{12} away from 0 (fig. 3e, f); in part, these changes reflect increases in size above nominal levels (fig. 1e, f).

Figures 4 and 5 show the biases (fig. 4a–c, 5a–c), coverage probabilities of 95% confidence intervals for β_1 (fig. 4d–f, 5d–f) and mean squared errors (MSE; fig. 4g–i, 5g–i) for true disease probabilities, $P(D = 1)$, of 10% (fig. 4) and 30% (fig. 5) when $\beta_1 = 0.25$. When the correct disease probability is used (fig. 4a, d, g, 5a, d, g), the estimates $\hat{\beta}_{1FM}$ and $\hat{\beta}_{1AW}$ are unbiased, have the same small MSE and are associated with nominal confidence interval coverage across the range of values of δ_{12} , both for $P(D = 1) = 10\%$ (fig. 4) and 30% (fig. 5). In contrast, $\hat{\beta}_{1RM}$ is severely biased and has sub-nominal confidence interval coverage and inflated MSE for values of δ_{12} away from 0. The good performance of the FM and AW procedures persists in the presence of misspecification of $P(D = 1)$ when $P(D = 1) = 10\%$ (fig. 4b, c, e, f, h, i). However, when $P(D = 1) = 30\%$, the FM and AW procedures yield slightly biased estimates of β_1 and slightly sub-nominal coverage of confidence intervals for large or small values of δ_{12} if $P(D = 1)$ is misspecified (fig. 5c, f).

Fig. 1. Type I error rates of association tests at the 5% nominal significance level. The true disease rates are 10% (a–c) and 30% (d–f).

Fig. 2. Simulation results for the estimated standard errors of $\hat{\beta}_{1RM}$ and $\hat{\beta}_{1RMU}$. The results are for $\beta_1 = 0$, $\delta_{12} = 0$ and the primary disease probability is 30%, based on 1,000 replicates. **a, b** Histograms of the estimated standard errors ($\hat{se}$) of $\hat{\beta}_{1RM}$ and $\hat{\beta}_{1RMU}$. **c, d** Scatter plots of $\hat{\beta}_{1RM}$ versus its estimated standard error

and $\hat{\beta}_{1RMU}$ versus estimated standard error, respectively. The red solid dots indicate the rejections of the null hypothesis for the Wald test with $\alpha = 0.05$. The type I error is 0.047 for the RM procedure and 0.113 for the RMU procedure.

Fig. 3. Powers of association tests at the alternative $\beta_1 = 0.25$. The true disease probabilities are 10% (a-c) and 30% (d-f).

Table 2. Type I error and power for a pre-selected SNP and for several values of δ_{12}

	Primary disease probability = 10%					Primary disease probability = 30%				
	$\delta_{12} = -0.5$	$\delta_{12} = -0.3$	$\delta_{12} = 0$	$\delta_{12} = 0.3$	$\delta_{12} = 0.5$	$\delta_{12} = -0.5$	$\delta_{12} = -0.3$	$\delta_{12} = 0$	$\delta_{12} = 0.3$	$\delta_{12} = 0.5$
$\beta_1 = 0$; size*										
FM	0.047	0.050	0.051	0.047	0.047	0.049	0.052	0.051	0.055	0.051
AW	0.051	0.060	0.045	0.058	0.051	0.049	0.052	0.051	0.056	0.052
RM	0.561	0.254	0.052	0.270	0.618	0.177	0.101	0.050	0.103	0.178
RMU	0.612	0.303	0.094	0.330	0.639	0.246	0.151	0.102	0.100	0.065
$\beta_1 = 0.25$; power*										
FM	0.511	0.521	0.515	0.528	0.526	0.669	0.671	0.662	0.681	0.673
AW	0.452	0.455	0.571	0.561	0.547	0.647	0.648	0.672	0.697	0.683
RM	0.080	0.255	0.750	0.978	0.999	0.338	0.504	0.729	0.896	0.957
RMU	0.127	0.316	0.783	0.919	0.904	0.441	0.568	0.669	0.468	0.218

Analyses use the correct primary disease probability.

* Based on 10,000 independent simulations for each column of four estimates.

For the additive genetic model, we fixed the minor allele frequency at $p = 0.4$, from which we have $P(G = 0) = 0.36$, $P(G = 1) = 0.48$ and $P(G = 2) = 0.16$. We set $\beta_0 = 0.1$ and let $\beta_1 = 0$ and 0.1 under the null and alternative hypotheses, respectively. For the disease model, we set $\delta_1 = \log(2)$ and $\delta_2 = \log(1.5)$, varied the value of δ_{12} from -0.5 to 0.5 and choose the value of δ_0 to yield a disease prevalence $P(D = 1)$ of 0.10 or 0.30 . The comparisons among the FM, RM and AW methods were very similar to those for dichotomous G with respect to estimation and testing for a pre-selected SNP (data not shown).

Genome-Wide Size and Power

In this section, we investigate the size and power of different methods to discover SNPs associated with the secondary phenotype in a GWAS. Assuming there were $N = 500,000$ independent SNP genotypes, we controlled the experiment-wise significance by setting $\alpha = 0.05/(5 \times 10^5) = 10^{-7}$. It may be reasonable to suppose that $\delta_{12} = 0$ for a large proportion of SNPs. As in Li et al. [6], we assume that 99% of SNPs have $\delta_{12} = 0$, and 1% of SNPs have δ_{12} independently distributed as $N(0, (\log(2)/2)^2)$, which implies that about 95% of non-zero δ_{12} values are in the interval $[-\log(2), \log(2)]$. We analytically evaluated the genome-wide type I error and power ($\beta_1 = 0$ and 0.25 for dichotomous G and $\beta_1 = 0$ and 0.1 for the additive genetic model) of Wald tests by averaging over the mixture distribution of δ_{12} (see [6]). Briefly, we computed the conditional power given δ_{12} from the non-central χ^2 distribution with non-centrality determined by $\{\beta_0, \beta_0, \delta_1, \delta_{12},$

PG_1 (or p), $P(D = 1)\}$ with the same parameter values as in the simulations for pre-selected SNPs. We then averaged this conditional power over the distribution of δ_{12} .

For dichotomous G , the genome-wide type I error and power for different methods are presented in table 3 for numbers of cases (n_1) and controls (n_0) each equal to 1,000, 5,000 or 10,000. The genome-wide type I error is always above the nominal level of 0.05 and usually equals 1.0 for the RM procedure, indicating that this test should not be used even if only a small proportion of SNPs have $\delta_{12} \neq 0$. Both FM and AW procedures have near nominal genome-wide type I error. However, for sample sizes $n_1 = n_0 = 5,000$ or $10,000$, the power of AW greatly exceeds that of FM when the true primary disease rate is 10% . For example, when $n_1 = n_0 = 5,000$, the power of AW is 70% , while that of FM is only 20% . Thus, substantial power gains can be achieved with the AW procedure. This power difference is smaller, but not negligible, when the primary disease probability is 30% (table 3). Misspecification of the primary disease probability has little effect on the size and power of the AW and FM procedures.

Qualitatively similar conclusions hold for the additive genetic model (table 4), which also shows that FM and WL perform very similarly. In particular, the size of RM greatly exceeds nominal levels and should be avoided. AW, FM and WL have nominal size, but AW is more powerful than FM or WL, especially for $P(D = 1) = 10\%$, but also for $P(D = 1) = 30\%$. Data in table 4 also indicate that these conclusions are robust to moderate misspecification of $P(D = 1)$, where the relative error is around 10% .

Fig. 4. Bias, coverage probabilities (CP) of 95% confidence intervals and MSE for different estimators with assumed primary disease probabilities $\xi_1 = 10, 11$ and 12% . The true primary disease probability is 10% and $\beta_1 = 0.25$.

Discussion

In this paper, we studied analyses of secondary dichotomous phenotypes for data from a case-control study of a primary disease that is not rare. We treated both dichotomous (dominant or recessive) and additive genetic

models. Using external information on the probability of the primary disease in the population, we developed an adaptive procedure, AW, and compared it with the previously published FM, WL, RM and RMU methods to analyze secondary phenotypes. We found that, when the primary disease is fairly common (10%), AW has good prop-

Fig. 5. Bias, coverage probabilities (CP) of 95% confidence intervals and MSE for different estimators with assumed primary disease probabilities $\xi_1 = 30, 33$ and 36% . The true primary disease probability is 30% and $\beta_1 = 0.25$.

erties both for inference on a pre-specified SNP and for discovery of SNP associations in a GWAS. In particular, it is more powerful in a GWAS than its chief competitors FM and WL, which perform similarly. AM performs well whether or not an interaction δ_{12} between the secondary phenotype and the SNP genotype affects the risk of the

primary disease. AW is also robust to misspecification of the probability of the primary disease in the general population, except when $P(D = 1)$ is large and $|\delta_{12}|$ is large. When the primary disease is common (30%), the performance of AW is similar to that of FM and WL, although AW still has a slight power advantage in GWAS. With a

Table 3. Genome-wide type I error and power with genome-wide significance level of 0.05 when δ_{12} is from a mixture distribution^a for the binary genetic model

$n_0 = n_1^c$	True primary disease probability = 10%						True primary disease probability = 30%					
	$\beta_1 = 0$; genome-wide size ^b			$\beta_1 = 0.25$; power ^b			$\beta_1 = 0$; genome-wide size ^b			$\beta_1 = 0.25$; power ^b		
	1,000	5,000	10,000	1,000	5,000	10,000	1,000	5,000	10,000	1,000	5,000	10,000
	Assumed primary disease probability = 10%						Assumed primary disease probability = 30%					
RM	1.000	1.000	1.000	0.004	0.703	0.996	0.104	1.000	1.000	0.003	0.677	0.997
FM	0.049	0.049	0.049	0.000	0.199	0.844	0.049	0.049	0.049	0.002	0.511	0.988
AW	0.049	0.049	0.049	0.003	0.699	0.997	0.049	0.049	0.049	0.003	0.676	0.998
	Assumed primary disease probability = 11%						Assumed primary disease probability = 32%					
RM	1.000	1.000	1.000	0.004	0.703	0.996	0.109	1.000	1.000	0.003	0.677	0.997
FM	0.049	0.049	0.049	0.000	0.206	0.853	0.049	0.049	0.049	0.002	0.541	0.991
AW	0.049	0.049	0.049	0.003	0.700	0.997	0.049	0.049	0.049	0.003	0.677	0.998
	Assumed primary disease probability = 12%						Assumed primary disease probability = 34%					
RM	1.000	1.000	1.000	0.004	0.703	0.996	0.110	1.000	1.000	0.003	0.677	0.997
FM	0.049	0.049	0.049	0.000	0.214	0.861	0.049	0.049	0.054	0.002	0.569	0.993
AW	0.049	0.049	0.049	0.003	0.700	0.997	0.049	0.049	0.055	0.003	0.677	0.998

^a The δ_{12} comes from a mixture distribution. With probability 0.99, $\delta_{12} = 0$. With probability 0.01, δ_{12} has a normal distribution with mean 0 and variance $(\log(2)/2)^2$.

^b Based on 10,000 independent simulations for each column of three estimates.

^c There are n_1 cases and n_0 controls in the case-control study.

Table 4. Genome-wide type I error and power with genome-wide significance level of 0.05 when δ_{12} is from a mixture distribution^a for the additive genetic model

$n_0 = n_1^c$	True primary disease probability = 10%						True primary disease probability = 30%					
	$\beta_1 = 0$; genome-wide size ^b			$\beta_1 = 0.15$; power ^b			$\beta_1 = 0$; genome-wide size ^b			$\beta_1 = 0.15$; power ^b		
	1,000	5,000	10,000	1,000	5,000	10,000	1,000	5,000	10,000	1,000	5,000	10,000
	Assumed primary disease probability = 10%						Assumed primary disease probability = 30%					
RM	1.000	1.000	1.000	0.002	0.429	0.971	0.527	1.000	1.000	0.001	0.442	0.975
FM	0.049	0.049	0.049	0.000	0.103	0.661	0.049	0.049	0.049	0.001	0.312	0.935
WL	0.049	0.049	0.049	0.000	0.100	0.653	0.049	0.049	0.049	0.001	0.305	0.931
AW	0.049	0.049	0.049	0.001	0.425	0.971	0.049	0.049	0.049	0.001	0.440	0.977
	Assumed primary disease probability = 11%						Assumed primary disease probability = 32%					
RM	1.000	1.000	1.000	0.002	0.461	0.977	0.701	1.000	1.000	0.002	0.466	0.979
FM	0.049	0.052	0.055	0.000	0.122	0.712	0.049	0.052	0.056	0.001	0.356	0.954
WL	0.049	0.051	0.054	0.000	0.119	0.705	0.049	0.052	0.056	0.001	0.351	0.952
AW	0.050	0.053	0.058	0.001	0.457	0.978	0.050	0.052	0.057	0.002	0.465	0.981
	Assumed primary disease probability = 12%						Assumed primary disease probability = 34%					
RM	1.000	1.000	1.000	0.002	0.492	0.982	0.996	1.000	1.000	0.002	0.488	0.982
FM	0.051	0.062	0.083	0.000	0.145	0.760	0.051	0.067	0.130	0.001	0.399	0.967
WL	0.051	0.060	0.077	0.000	0.141	0.752	0.051	0.069	0.139	0.001	0.395	0.966
AW	0.052	0.068	0.097	0.002	0.488	0.983	0.051	0.069	0.132	0.002	0.488	0.984

^a The δ_{12} comes from a mixture distribution. With probability 0.99, $\delta_{12} = 0$. With probability 0.01, δ_{12} has a normal distribution with mean 0 and variance $(\log(2)/2)^2$.

^b Based on 10,000 independent simulations for each column of three estimates.

^c There are n_1 cases and n_0 controls in the case-control study.

common primary disease, one may prefer to use WL because it is easy to compute and has previously been used and studied. For dichotomous G , we showed theoretically that the weighted logistic methods proposed by Richardson et al. [3] and Monsees et al. [4] are equivalent to maximum likelihood, which we call FM, for a primary disease model in which there is no restriction on δ_{12} but in which the probability of the primary disease is known. For the additive genetic model, WL yields numerical estimates close, but not identical, to FM.

The maximum likelihood procedure that assumes $\delta_{12} = 0$ and that requires external information on the probability of the primary disease, which we call RM, is the most efficient method when $\delta_{12} = 0$, but it does not control size when $\delta_{12} \neq 0$ and can be very misleading. A maximum likelihood procedure that assumes $\delta_{12} = 0$ and that does not require external information on the probability of the primary disease [2], which we call RMU, does not control size even when $\delta_{12} = 0$ and should be avoided. The problem arises because under the null hypothesis $\beta_1 = 0$ and assuming $\delta_{12} = 0$, other parameter estimates are biased and unstable, leading to incorrect estimates of the information matrix and underestimates of the standard error of $\hat{\beta}_{1RMU}$. Lin and Zeng [2] alluded to this instability, and their publically available software at <http://www.bios.unc.edu/~lin/software/SPREG/> requires knowing $P(D = 1)$. Thus, for a dichotomous secondary phenotype, there seems to be no practical alternative to obtaining information on the probability of the primary disease and using it in procedures like AW, WL and FM. When there is substantial uncertainty concerning the probability of the primary disease, sensitivity analyses may be helpful. It is encouraging, however, that AW and FM are rather robust to misspecification of this probability for the dichotomous genetic model and are robust to the moderate misspecification (10% relative error) of this probability for the additive genetic model.

If case-control data for a common primary disease arise in a nested case-control study within a cohort, the cohort information will provide a direct estimate of the probability of the primary disease. Our unreported simulations demonstrate that very similar inference for the secondary endpoint is obtained from AW and FM procedures whether one assumes that the probability of the primary disease is known or whether one instead inserts the estimate from the cohort data.

Wang and Shete [5] used a series of estimating equations to solve for the parameters in the reduced model, which in our notation are $\theta = \{\beta_0, \beta_1, \delta_0, \delta_1, \delta_2, PG_0\}$. Their assumptions include: (1) $\delta_{12} = 0$, and (2) both the proba-

bility of the primary disease, $P(D = 1)$, and the probability of the secondary phenotype, $P(X = 1)$, are known. More work is needed to determine how robust their procedures are to violations of these assumptions and to assess their efficiency compared to maximum likelihood procedures. However, it is likely that violation of the assumption $\delta_{12} = 0$ can induce misleading results, as for the maximum likelihood procedures RM and RMU.

Several topics warrant additional research, including extensions of the models to allow for covariates and for continuous secondary phenotypes. If only a few covariates are needed, the methods of the present paper can be used within strata defined by the covariates.

Acknowledgements

Part of Dr. Li's research and Dr. Gail's research were supported by the Intramural Research Program of the Division of Cancer Epidemiology and Genetics of the National Cancer Institute. This research has utilized the high-performance computational capabilities of the Biowulf PC/Linux cluster at the National Institutes of Health, Bethesda, Md., USA, and at the Center for Health Informatics and Bioinformatics at New York University Langone Medical Center, New York, N.Y., USA.

Appendices

Derivation of $\hat{\beta}_{IFM}$

Under equations (1) and (2), we need to estimate seven parameters: $\theta_{FM} = \{\beta_0, \beta_1, \delta_0, \delta_1, \delta_2, \delta_{12}, PG_0\}$ by maximizing the likelihood (7). In the Appendices, we denote $P(D = 1) \equiv \xi_1$. In order to avoid complicated calculations, we take the logarithm of equation (7) and change the variables first. By letting

$$p_{ijk} = P(D = i, G = j, X = k) = \frac{\exp(i(\delta_0 + \delta_1 j + \delta_2 k + \delta_{12} jk)) \exp(k(\beta_0 + \beta_1 j))}{1 + \exp(\delta_0 + \delta_1 j + \delta_2 k + \delta_{12} jk) + \exp(\beta_0 + \beta_1 j)} PG_j; \quad i, j, k = 0, 1, \quad (A1)$$

and with the constraints

$$\sum_{i,j,k=0}^1 p_{ijk} = 1 \quad \text{and} \quad \sum_{j,k=0}^1 p_{1jk} = \xi_1,$$

we have the equivalent log-likelihood

$$\begin{aligned} \tilde{l}_{FM} = & r_{000} \log(1 - \xi_1 - p_{001} - p_{010} - p_{011}) + r_{010} \log(p_{010}) \\ & + r_{001} \log(p_{001}) + r_{011} \log(p_{011}) \\ & + r_{100} \log(p_{100}) + r_{110} \log(p_{110}) + r_{101} \log(p_{101}) \\ & + r_{111} \log(\xi_1 - p_{100} - p_{110} - p_{101}) \end{aligned} \quad (A2)$$

which has six parameters $\varphi_{FM} = (p_{010}, p_{001}, p_{011}, p_{100}, p_{110}, p_{101})$. Using the standard maximization technique, we have the following estimates:

$$\hat{p}_{001} = \frac{(1 - \xi_1)r_{001}}{n_0}, \hat{p}_{010} = \frac{(1 - \xi_1)r_{010}}{n_0}, \hat{p}_{011} = \frac{(1 - \xi_1)r_{011}}{n_0},$$

$$\hat{p}_{100} = \frac{\xi_1 r_{100}}{n_1}, \hat{p}_{101} = \frac{\xi_1 r_{101}}{n_1}, \text{ and } \hat{p}_{110} = \frac{\xi_1 r_{110}}{n_1}.$$

Solving for the original parameters using equation (A1), we obtain

$$\hat{\delta}_{0FM} = \log\left(\frac{\hat{p}_{100}}{\hat{p}_{000}}\right) = \log\left(\frac{\xi_1}{1 - \xi_1} \frac{r_{100}}{r_{000}} \frac{n_0}{n_1}\right),$$

$$\hat{\delta}_{1FM} = \log\left(\frac{\hat{p}_{110}}{\hat{p}_{010}} \frac{\hat{p}_{000}}{\hat{p}_{100}}\right) = \log\left(\frac{r_{110}}{r_{010}} \frac{r_{000}}{r_{100}}\right),$$

$$\hat{\delta}_{2FM} = \log\left(\frac{\hat{p}_{101}}{\hat{p}_{001}} \frac{\hat{p}_{000}}{\hat{p}_{100}}\right) = \log\left(\frac{r_{101}}{r_{001}} \frac{r_{000}}{r_{100}}\right),$$

$$\hat{\delta}_{12FM} = \log\left(\frac{\hat{p}_{111}}{\hat{p}_{011}} \frac{\hat{p}_{010}}{\hat{p}_{110}} \frac{\hat{p}_{001}}{\hat{p}_{101}} \frac{\hat{p}_{100}}{\hat{p}_{000}}\right) = \log\left(\frac{r_{111}}{r_{011}} \frac{r_{010}}{r_{110}} \frac{r_{001}}{r_{101}} \frac{r_{100}}{r_{000}}\right),$$

$$P\hat{G}_{0FM} = \hat{p}_{100} + \hat{p}_{101} + \hat{p}_{000} + \hat{p}_{001} = \frac{\xi_1(r_{100} + r_{101})}{n_1} + \frac{\xi_0(r_{000} + r_{001})}{n_0},$$

$$\hat{\beta}_{0FM} = \log\left(\frac{\hat{p}_{101} + \hat{p}_{001}}{\hat{p}_{100} + \hat{p}_{000}}\right) = \log\left(\frac{\xi_1 r_{101} n_0 + (1 - \xi_1) r_{001} n_1}{\xi_1 r_{100} n_0 + (1 - \xi_1) r_{000} n_1}\right),$$

and

$$\hat{\beta}_{1FM} = \log\left(\frac{(\hat{p}_{111} + \hat{p}_{011})(\hat{p}_{100} + \hat{p}_{000})}{(\hat{p}_{110} + \hat{p}_{010})(\hat{p}_{101} + \hat{p}_{001})}\right)$$

$$= \log\left(\frac{\xi_1 r_{111} n_0 + (1 - \xi_1) r_{011} n_1}{\xi_1 r_{110} n_0 + (1 - \xi_1) r_{010} n_1} \cdot \frac{\xi_1 r_{100} n_0 + (1 - \xi_1) r_{000} n_1}{\xi_1 r_{101} n_0 + (1 - \xi_1) r_{001} n_1}\right),$$

where

$$\hat{p}_{111} = \xi_1 - \hat{p}_{100} - \hat{p}_{110} - \hat{p}_{101} = \frac{\xi_1 r_{111}}{n_1}.$$

$$\hat{p}_{000} = 1 - \xi_1 - \hat{p}_{001} - \hat{p}_{010} - \hat{p}_{011} = \frac{(1 - \xi_1) r_{000}}{n_0}$$

Proof of $\hat{\beta}_{1FM} = \hat{\beta}_{1W}$ with

$$w_i = P(D = i) / n_i = \xi_i / n_i,$$

$$\hat{\beta}_{1W} = \log\left(\frac{w_0 r_{011} + w_1 r_{111}}{w_0 r_{010} + w_1 r_{110}} \cdot \frac{w_0 r_{000} + w_1 r_{100}}{w_0 r_{001} + w_1 r_{101}}\right) = \hat{\beta}_{1FM}.$$

Derivation of $\hat{\beta}_{IRM}$ with Known Disease Rate when $\delta_{12} = 0$

Assuming $\delta_{12} = 0$, we have the reduced disease model logit $(P(D = 1 | G, X)) = \delta_0 + \delta_1 G + \delta_2 X$. Including the parameters of equation (1), we need to estimate six parameters $\theta_{RM} = \{\beta_0, \beta_1, \delta_0, \delta_1, \delta_2, PG_0\}$. To do this, we maximize the likelihood (6). As in equation (A1), we re-parameterize first. By letting $p_{ijk} = P(D = i, X = j, G = k); i, j, k = 0, 1$ and incorporating the constraints

$$\sum_{i,j,k=0}^1 p_{ijk} = 1 \quad \text{and} \quad \sum_{j,k=0}^1 p_{1jk} = \xi_1,$$

we obtain

$$\tilde{l}_{RM} = r_{000} \log(1 - \xi_1 - p_{001} - p_{010} - p_{011}) + r_{010} \log(p_{010})$$

$$+ r_{001} \log(p_{001}) + r_{011} \log(p_{011}) + r_{100} \log(p_{100}) + r_{110} \log(p_{110})$$

$$+ r_{101} \log\left(\frac{p_{001} p_{010} p_{100} (\xi_1 - p_{100} - p_{110})}{p_{110} p_{011} (1 - \xi_1 - p_{001} - p_{010} - p_{011}) + p_{001} p_{010} p_{100}}\right)$$

$$+ r_{111} \log\left(\frac{p_{110} p_{011} (1 - \xi_1 - p_{001} - p_{010} - p_{011}) (\xi_1 - p_{100} - p_{110})}{p_{110} p_{011} (1 - \xi_1 - p_{001} - p_{010} - p_{011}) + p_{001} p_{010} p_{100}}\right) \quad (A3)$$

which has only five parameters $\varphi_{RM} = (p_{010}, p_{001}, p_{011}, p_{100}, p_{110})$. Using the SAS procedure PROC NLPTR, we obtained the maximum likelihood estimators $\hat{\varphi}_{RM}$. Re-expressing our results in terms of the original parameters, we have

$$\hat{\delta}_{0RM} = \log\left(\frac{\hat{p}_{100}}{\hat{p}_{000}}\right), \hat{\delta}_{1RM} = \log\left(\frac{\hat{p}_{110}}{\hat{p}_{010}} \frac{\hat{p}_{000}}{\hat{p}_{100}}\right), \hat{\delta}_{2RM} = \log\left(\frac{\hat{p}_{101}}{\hat{p}_{001}} \frac{\hat{p}_{000}}{\hat{p}_{100}}\right),$$

$$P\hat{G}_{0RM} = \hat{p}_{100} + \hat{p}_{101} + \hat{p}_{000} + \hat{p}_{001}, \hat{\beta}_{0RM} = \log\left(\frac{\hat{p}_{101} + \hat{p}_{001}}{\hat{p}_{100} + \hat{p}_{000}}\right),$$

$$\hat{\beta}_{1RM} = \log\left(\frac{(\hat{p}_{111} + \hat{p}_{011})(\hat{p}_{100} + \hat{p}_{000})}{(\hat{p}_{110} + \hat{p}_{010})(\hat{p}_{101} + \hat{p}_{001})}\right),$$

where

$$\hat{p}_{000} = 1 - \xi_1 - \hat{p}_{001} - \hat{p}_{010} - \hat{p}_{011},$$

$$\hat{p}_{101} = \frac{\hat{p}_{001} \hat{p}_{010} \hat{p}_{100} (\xi_1 - \hat{p}_{100} - \hat{p}_{110})}{\hat{p}_{110} \hat{p}_{011} (1 - \xi_1 - \hat{p}_{001} - \hat{p}_{010} - \hat{p}_{011}) + \hat{p}_{001} \hat{p}_{010} \hat{p}_{100}},$$

and

$$\hat{p}_{111} = \frac{\hat{p}_{110} \hat{p}_{011} (1 - \xi_1 - \hat{p}_{001} - \hat{p}_{010} - \hat{p}_{011}) (\xi_1 - \hat{p}_{100} - \hat{p}_{110})}{\hat{p}_{110} \hat{p}_{011} (1 - \xi_1 - \hat{p}_{001} - \hat{p}_{010} - \hat{p}_{011}) + \hat{p}_{001} \hat{p}_{010} \hat{p}_{100}}.$$

Derivation of $\hat{\beta}_{IRMU}$ with Unknown Disease Rate when $\delta_{12} = 0$

As in Lin and Zeng [2], we can obtain the estimate of β_1 without knowing the disease rate. Using the general retrospective likelihood (6), we express the log-likelihood with a new set of parameters $p_{ijk} = P(D = i, X = j, G = k); i, j, k = 0, 1$ as follows:

$$\tilde{l}_{RMU} = r_{000} \log(p_{000}) + r_{010} \log(p_{010}) + r_{001} \log(p_{001})$$

$$+ r_{011} \log(p_{011}) + r_{100} \log(p_{100}) + r_{110} \log(p_{110})$$

$$+ r_{101} \log\left(\frac{p_{001} p_{010} p_{100} (\xi_1 - p_{100} - p_{110})}{p_{110} p_{011} p_{000} + p_{001} p_{010} p_{100}}\right)$$

$$+ r_{111} \log\left(\frac{p_{110} p_{011} p_{000} (\xi_1 - p_{100} - p_{110})}{p_{110} p_{011} p_{000} + p_{001} p_{010} p_{100}}\right)$$

$$- n_0 \log(p_{000} + p_{001} + p_{010} + p_{011})$$

which has only six parameters $\varphi_{RMU} = (p_{000}, p_{010}, p_{001}, p_{011}, p_{100}, p_{110})$. Using the SAS procedure PROC NLPTR, we obtained the maximum likelihood estimators $\hat{\varphi}_{RMU}$. In terms of the original parameters, we have

$$\begin{aligned}\hat{\delta}_{0RMU} &= \log\left(\frac{\hat{p}_{100}}{\hat{p}_{000}}\right), \quad \hat{\delta}_{1RMU} = \log\left(\frac{\hat{p}_{110}}{\hat{p}_{010}} \frac{\hat{p}_{000}}{\hat{p}_{100}}\right), \\ \hat{\delta}_{2RMU} &= \log\left(\frac{\hat{p}_{101}}{\hat{p}_{001}} \frac{\hat{p}_{000}}{\hat{p}_{100}}\right), \quad P\hat{G}_{0RMU} = \hat{p}_{100} + \hat{p}_{101} + \hat{p}_{000} + \hat{p}_{001}, \\ \hat{\beta}_{0RMU} &= \log\left(\frac{\hat{p}_{101} + \hat{p}_{001}}{\hat{p}_{100} + \hat{p}_{000}}\right), \\ \text{and } \hat{\beta}_{1RMU} &= \log\left(\frac{(\hat{p}_{111} + \hat{p}_{011})(\hat{p}_{100} + \hat{p}_{000})}{(\hat{p}_{110} + \hat{p}_{010})(\hat{p}_{101} + \hat{p}_{001})}\right),\end{aligned}$$

where

$$\hat{p}_{101} = \frac{\hat{p}_{001}\hat{p}_{010}\hat{p}_{100}(\xi_1 - \hat{p}_{100} - \hat{p}_{110})}{\hat{p}_{110}\hat{p}_{011}(1 - \xi_1 - \hat{p}_{001} - \hat{p}_{010} - \hat{p}_{011}) + \hat{p}_{001}\hat{p}_{010}\hat{p}_{100}}$$

and

$$\hat{p}_{111} = \frac{\hat{p}_{110}\hat{p}_{011}(1 - \xi_1 - \hat{p}_{001} - \hat{p}_{010} - \hat{p}_{011})(\xi_1 - \hat{p}_{100} - \hat{p}_{110})}{\hat{p}_{110}\hat{p}_{011}(1 - \xi_1 - \hat{p}_{001} - \hat{p}_{010} - \hat{p}_{011}) + \hat{p}_{001}\hat{p}_{010}\hat{p}_{100}}.$$

Variance Estimator of $\hat{\beta}_{1AW}$

Ignoring the variability of the variance estimator of $\hat{\beta}_{1FM}$: $\hat{\sigma}_{FM}$, we identify the adaptively weighted estimator as a function of $(\hat{\beta}_{1FM}, \hat{\delta}_{12FM}, \hat{\beta}_{1RM})$ [7] and we have

$$\begin{aligned}\hat{\beta}_{1AW} &= g(\hat{\beta}_{1FM}, \hat{\delta}_{12FM}, \hat{\beta}_{1RM}) \\ &= \frac{\hat{\delta}_{12FM}^2}{\hat{\delta}_{12FM}^2 + \hat{\sigma}_{FM}^2} \hat{\beta}_{1FM} + \frac{\hat{\sigma}_{FM}^2}{\hat{\delta}_{12FM}^2 + \hat{\sigma}_{FM}^2} \hat{\beta}_{1RM}.\end{aligned}$$

In order to derive the variance estimator of $\hat{\beta}_{1AW}$, we first obtain the joint asymptotic distribution of the MLEs $(\hat{\beta}_{1FM}, \hat{\delta}_{12FM}, \hat{\beta}_{1RM})$. $\hat{\varphi}_{FM}$ and $\hat{\varphi}_{RM}$ are MLEs of equations (A2) and (A3) using the full disease model and the reduced disease model, respectively.

ly. Let $I_{\varphi_{FM}}$ (6×6) and $I_{\varphi_{RM}}$ (5×5) denote the observed information matrices for likelihood with the full (A2) and reduced (A3) model, respectively. Then, asymptotically,

$$\begin{pmatrix} \hat{\beta}_{1FM} - \beta_{1F} \\ \hat{\delta}_{12FM} - \delta_{12F} \\ \hat{\beta}_{1RM} - \beta_{1R} \end{pmatrix} = \begin{pmatrix} -(\nabla \hat{\beta}_{1FM}(\hat{\varphi}_{FM}))^T - I_{\varphi_{FM}}^{-1} U_{\varphi_{FM}} + O_p(n^{-1/2}) \\ -(\nabla \hat{\delta}_{12FM}(\hat{\varphi}_{FM}))^T I_{\varphi_{FM}}^{-1} U_{\varphi_{FM}} + O_p(n^{-1/2}) \\ -(\nabla \hat{\beta}_{1RM}(\hat{\varphi}_{RM}))^T I_{\varphi_{RM}}^{-1} U_{\varphi_{RM}} + O_p(n^{-1/2}) \end{pmatrix}$$

where $U_{\varphi_{FM}}$ and $U_{\varphi_{RM}}$ denote the score functions for φ_{FM} and φ_{RM} , respectively, and n denotes the total sample size. $\nabla \hat{\beta}_{1FM}(\hat{\varphi}_{FM})$, $\nabla \hat{\delta}_{12FM}(\hat{\varphi}_{FM})$ and $\nabla \hat{\beta}_{1RM}(\hat{\varphi}_{RM})$ are the gradient matrices of $\hat{\beta}_{1FM}$, $\hat{\delta}_{12FM}$ and $\hat{\beta}_{1RM}$ with respect to φ_{FM} and φ_{RM} , evaluated at $\hat{\varphi}_{FM}$ and $\hat{\varphi}_{RM}$.

Then, using the delta method, we have the variance estimator of $\hat{\beta}_{1AW}$ given by

$$\hat{\sigma}_{AM}^2 = \{g'(\hat{\beta}_{1FM}, \hat{\delta}_{12FM}, \hat{\beta}_{1RM})\}^T \Sigma_{ML} g'(\hat{\beta}_{1FM}, \hat{\delta}_{12FM}, \hat{\beta}_{1RM}),$$

where g' is the gradient matrix of g (A4) with respect to $(\hat{\beta}_{1FM}, \hat{\delta}_{12FM}, \hat{\beta}_{1RM})$ and Σ_{ML} is the asymptotic variance-covariance matrix of $(\hat{\beta}_{1FM}, \hat{\delta}_{12FM}, \hat{\beta}_{1RM})$. We partition Σ_{ML} as

$$\Sigma_{ML} = \begin{pmatrix} \sum_{12 \cdot 12} & \sum_{12 \cdot 3} \\ \sum_{3 \cdot 12} & \sum_{33} \end{pmatrix}$$

where

$$\begin{aligned}\sum_{12 \cdot 12} &= (\nabla \hat{\beta}_{1FM}(\hat{\varphi}_{FM}), \nabla \hat{\delta}_{12FM}(\hat{\varphi}_{FM}))^T \\ &\quad I_{\varphi_{FM}}^{-1} (\nabla \hat{\beta}_{1FM}(\hat{\varphi}_{FM}), \nabla \hat{\delta}_{12FM}(\hat{\varphi}_{FM})), \\ \sum_{33} &= (\nabla \hat{\beta}_{1RM}(\hat{\varphi}_{RM}))^T I_{\varphi_{RM}}^{-1} \nabla \hat{\beta}_{1RM}(\hat{\varphi}_{RM}),\end{aligned}$$

and

$$\begin{aligned}\sum_{12 \cdot 3} &= \sum_{3 \cdot 12} \\ &= \begin{pmatrix} (\nabla \hat{\beta}_{1FM}(\hat{\varphi}_{FM}))^T I_{\varphi_{FM}}^{-1} E(U_{\varphi_{FM}} U_{\varphi_{RM}}^T) I_{\varphi_{RM}}^{-1} \nabla \hat{\beta}_{1RM}(\hat{\varphi}_{RM}) \\ (\nabla \hat{\delta}_{12FM}(\hat{\varphi}_{FM}))^T I_{\varphi_{FM}}^{-1} E(U_{\varphi_{FM}} U_{\varphi_{RM}}^T) I_{\varphi_{RM}}^{-1} \nabla \hat{\beta}_{1RM}(\hat{\varphi}_{RM}) \end{pmatrix}.\end{aligned}$$

References

- Jiang Y, Scott AJ, Wild CJ: Secondary analysis of case-control data. *Stat Med* 2006;25: 1323–1339.
- Lin DY, Zeng D: Proper analysis of secondary phenotype data in case-control association studies. *Genet Epidemiol* 2009;33:256–265.
- Richardson DB, Rzehak P, Klenk J, Weiland SK: Analyses of case-control data for additional outcomes. *Epidemiology* 2007;18: 441–445.
- Monsees GM, Tamimi RM, Kraft P: Genome-wide association scans for secondary traits using case-control samples. *Genet Epidemiol* 2009;33:717–728.
- Wang J, Shete S: Estimation of odds ratios of genetic variants for the secondary phenotypes associated with primary diseases. *Genet Epidemiol* 2011;35:190–200.
- Li H, Gail MH, Berndt S, Chatterjee N: Using cases to strengthen inference on the association between single nucleotide polymorphisms and a secondary phenotype in genome-wide association studies. *Genet Epidemiol* 2010;34:427–433.
- Mukherjee B, Chatterjee N: Exploiting gene-environment independence for analysis of case-control studies: an empirical Bayes-type shrinkage estimator to trade-off between bias and efficiency. *Biometrics* 2008; 64:685–694.