Published in final edited form as:

Hum Hered. 2015 ; 80(4): 187–195. doi:10.1159/000446957.

Retrospective Association Analysis of Binary Traits: Overcoming Some Limitations of the Additive Polygenic Model

Duo Jiang[1], **Joelle Mbatchou**[2], and **Mary Sara McPeek**[2,3]

[1]Department of Statistics, Oregon State University, Corvallis, OR 97331, USA

[2]Departments of Statistics, University of Chicago, Chicago, IL 60637, USA

[3]Departments of Human Genetics, University of Chicago, Chicago, IL 60637, USA

Abstract

Case-control genetic association analysis is an extremely common tool in human complex trait mapping. From a statistical point of view, analysis of binary traits poses somewhat different challenges from analysis of quantitative traits. Desirable features of a binary trait mapping approach would include (1) phenotype modeled as binary, with appropriate dependence between the mean and variance; (2) appropriate correction for relevant covariates; (3) appropriate correction for sample structure of various types, including related individuals, admixture, and other types of population structure; (4) both fast and accurate computations; (5) robustness to ascertainment and other types of phenotype model misspecification; (6) ability to leverage partially missing data to increase power. We review these challenges and argue, both theoretically and in simulations, for the value of retrospective association analysis as a way to overcome some of the limitations of the phenotype model, including model misspecification due to ascertainment. We describe two recent retrospective methods, CARAT and CERAMIC, that are designed to meet criteria (1)-(6).

Keywords

binary traits; retrospective analysis; case-control association analysis; ascertainment; mixed models; population structure; pedigree; quasi-likelihood; CARAT; CERAMIC

Introduction

Many complex traits of interest come in the form of a binary variable (e.g., presence or absence of disease). Consequently, case-control genetic association analysis has an important role to play in elucidation of the genetic architecture of complex traits. Compared to analysis of continuous traits, analysis of binary traits presents particular modeling and computational challenges. Unlike a normal random variable, a binary random variable has its mean restricted to lie between 0 and 1 and has a fixed relationship between its mean and variance. When these features are combined with the need to account for covariates and for

Address for correspondence: Mary Sara McPeek, Department of Statistics, University of Chicago, 5747 S. Ellis Avenue, Jones, Chicago, IL 60637, Phone : (773) 702-7554 Fax : (773) 702-9810, mcpeek@uchicago.edu.

sample structure, by which we mean either population structure or family structure or both, in an association analysis, the resulting models can be computationally demanding or infeasible for large-scale analysis. In addition, issues of ascertainment arise even more commonly in case-control analysis than in continuous trait analysis. One concern about ascertainment is that it can be a major source of misspecification of the trait model in the ascertained sample. In general such model misspecification can lead to compromised type 1 error control and/or power loss.

In recent literature[1–3], it is common to suggest that an additive polygenic model on the linear scale, which is a type of linear mixed model (LMM), be used for binary trait analysis, ignoring the mean and variance restrictions on the binary trait. However, based on recent work[4, 5], we argue that this can result in a loss of power due to ignoring the special mean-variance structure of a binary trait. We also explore the advantages of a retrospective association analysis (in which the genotype at a tested variant is treated as random, and the analysis is performed conditional on the phenotype and covariates) compared to the more common prospective association analysis (in which the phenotype is treated as random, and the analysis is performed conditional on the genotype and covariates). We find that, compared to prospective association analysis, retrospective association analysis is less sensitive to phenotype model misspecification and less susceptible to power loss due to phenotype and covariate-based ascertainment. This is an important consideration in light of the fact that the trait model is typically unknown and considering the possibly strong effects of ascertainment.

Finally, we give an overview of two recently-developed methods, CARAT and CERAMIC, which account for covariate effects and sample structure in a way that is tailored for binary traits and which use retrospective association analysis to achieve robustness and high power in the presence of phenotype model misspecification. CARAT focuses on adjustment for population structure, while CERAMIC handles samples with related individuals and is able to incorporate partially missing data.

Binary trait modeling with covariates and sample structure

Consider a sample of n individuals, and let $\boldsymbol{Y}$ be a vector of length n whose ith element, Y_i, is the value of the binary phenotype for individual i. Suppose we also observe $k-1 \geq 0$ non-trivial covariates, encoded in an $n \times k$ design matrix, $\boldsymbol{X}$, which always has a column of ones as its first column (representing an intercept term). Let $\boldsymbol{X}_i$ denote the ith row of $\boldsymbol{X}$, i.e., $\boldsymbol{X_i}$ contains the covariate values for individual i. Suppose association is to be tested with a biallelic variant that is encoded in the vector, $\boldsymbol{G}$, of length n, where G_i denotes the minor allele count (0, 1 or 2) of individual i at the tested variant.

A special feature of binary random variables is that the mean, μ, and variance, σ^2, of any given binary random variable are constrained by $\sigma^2 = \mu(1-\mu)$, $0 \leq \mu \leq 1$. Thus, in modeling the binary vector $\boldsymbol{Y}$ conditional on $\boldsymbol{X}$ and $\boldsymbol{G}$, it is natural to (at least, initially) consider generalized linear models (GLMs) of the form

$$Y_i|\boldsymbol{X},\boldsymbol{G},\sim\text{Bernoulli}(\mu_i),\ \text{independently},\ \text{with}\ g(\mu_i)=\boldsymbol{X}_i\boldsymbol{\beta}+G_i\gamma,\ i=1,\cdots,n, \quad (1)$$

where $\boldsymbol{\beta}$ is a k-dimensional vector of unknown covariate effects, γ is the unknown scalar association parameter, and g is a known link function. For example, logistic regression would correspond to the choice of the logit link function for g, while a liability threshold model would correspond to the choice of the probit link function for g.

When the binary phenotype $\boldsymbol{Y}$ represents a complex trait, we would typically expect it to be influenced by additional genetic loci beyond the tested variant $\boldsymbol{G}$. When there is also relatedness or population structure among the sampled individuals, then the effects of additional genetic loci would typically lead to misspecification of the model in equation 1, unless these effects were taken into account somehow. If sample structure were known (or could be estimated from genome-wide data) and could be well-modeled by including a small number of fixed effects (e.g. population membership indicators or principal components), then these could be included in the covariate matrix $\boldsymbol{X}$. However, with many types of sample structure, including related individuals, admixture and other non-trivial types of population structure, such an approach can be inadequate.

For continuously-varying traits, a time-honored (and newly popular) approach to this problem is to use a LMM. For example, when additive polygenic effects and independent noise are modeled, the LMM could be written as

$$\boldsymbol{Y}=\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{G}\gamma+\boldsymbol{\varepsilon},\ \boldsymbol{\varepsilon}\sim N(0,\sigma_a^2\boldsymbol{\Phi}+\sigma_e^2\boldsymbol{I}), \quad (2)$$

where σ_a^2 and σ_e^2 are variance component parameters corresponding to additive polygenic effects and environmental errors, respectively, and where $\boldsymbol{\Phi}$ is an $n \times n$ genetic relationship matrix that quantifies the overall genetic similarity between individuals. For example, $\boldsymbol{\Phi}$ could be the kinship matrix[6] computed from a known pedigree, or it could be an empirical genetic relatedness matrix[1, 7] computed from genome-wide data. Association analysis of a continuous trait by testing the null hypothesis $\gamma = 0$ in the LMM of equation 2 has the advantage of working well (in terms of type 1 error and power) for a variety of types of population structure. It is frequently suggested that association analysis of binary traits be carried out in the same way[1–3].

However, use of a linear model for a binary variable has some obvious drawbacks, as has long been noted in the statistical literature. For exmple, Nerlove and Press[8] note that the resulting heteroscedasticity leads to “inefficient estimators and imprecise predictions” and that “the usual tests of significance for the estimated coefficients do not apply.” An obvious approach to combining the benefits of GLM and LMM models would be the use of a generalized linear mixed model (GLMM). For example, a model that includes both additive polygenic effects and independent noise (the latter in the form of Bernoulli error) is

$$Y_i|\boldsymbol{X},\boldsymbol{G},\varepsilon_i \sim \text{Bernoulli}(\mu_i),\ \text{independently},\ \text{with} g(\mu_i)=\boldsymbol{X}_i\boldsymbol{\beta}+G_i\gamma+\varepsilon_i,\ i=1,\cdots,n, \quad (3)$$

where $\boldsymbol{\varepsilon}=(\boldsymbol{\varepsilon}_i,\cdots,\varepsilon_n)^T \sim N(0,\sigma_a^2\boldsymbol{\Phi})$ and $\boldsymbol{g}$ is a known link function, for example, the logit link function.

The drawbacks of GLMM are largely computational. Estimating parameters in GLMMs involves intractable high-dimensional integration due to the presence of random effects. Thus, most implementations will rely on some level of approximation of this integral, with a tradeoff between speed and accuracy. These approaches can be broken down into two categories: frequentist/deterministic versus Bayesian/stochastic approaches.

A widely used frequentist approach is the Penalized Quasi-Likelihood method (PQL) [9–11] which, while being quite fast and flexible, will behave badly for binary data and tend to give estimates that are biased towards zero [12]. Despite the known problems, this method remains widely used for binary data [13–15] because of its speed. The Laplace Approximation [16], another frequentist approach, is slower than PQL but will give more accurate estimates. Finally, Gauss-Hermite Quadrature [17, 18] is another method which, while being slower than the previous two approaches, leads to even higher estimation accuracy. Bayesian approaches, on the other hand, will rely on Markov chain Monte Carlo (MCMC) algorithms [19, 20]. Hence, although accurate and highly flexible, these methods will be slower than the deterministic ones, all the more so for larger data sets. In addition, when using MCMC, one has to specify priors for the model parameters, which is not a trivial task [21], as well as perform checks for convergence [22].

Model misspecification

In genetic association studies, the underlying distribution of the trait variable and how it relates to the covariates is typically unknown, and it can be hard to verify whether an assumed model for the phenotype is correctly specified. Consequently, misspecification of the phenotypic model is not uncommon, especially for binary traits, and often goes undetected in genetic association testing. For example, this could happen if a linear model assuming normality is imposed on a binary phenotype, if an incorrect link function is used in GLMM, if important variance components are neglected (e.g. dominance polygenic variance), if quadratic effects of the covariates are inadequately accounted for by the model, or if sample structure is not appropriately modeled.

Another frequent source of model misspecification is non-random ascertainment. Because power to detect association with a binary trait is typically higher when the sample is approximately evenly-divided between case and control (all other things being equal), oversampling of cases is an extremely common practice in case-control association studies. The resulting distribution of (***X***, ***G***, ***Y***) in the ascertained sample differs from what it would be in a simple random sample. In most genetic association testing methods, ascertained data are analyzed using a population-based model in which ascertainment is ignored due to the difficulty of modeling it. An exception is a method that explicitly incorporates sampling

probabilities when detailed information about the sampling criterion is available[23]. Two recently proposed methods[24, 25] analyze ascertained case-control data using modified population-based models that incorporate known prevalence of the disease and a prevalence-adjusted heritability estimate. In genetic association analysis, the use of a population-based model in which the ascertainment scheme is ignored can result in the model being misspecified for ascertained data.

Prospective vs. retrospective testing To assess the statistical significance of a genetic association test, most current methods take either a prospective or a retrospective approach. In the prospective approach, the phenotype of interest, $\boldsymbol{Y}$, is taken as random, and its distribution is modeled conditional on the genotype and covariates, ($\boldsymbol{X}$, $\boldsymbol{G}$). Examples of methods using this approach include GLOGS[26] and GMMAT[15], which are based on a prospective GLMM, and EMMAX[1], which is based on a LMM. Alternatively, assessment of significance can be done retrospectively by considering the conditional distribution of the genotype at the variant of interest, $\boldsymbol{G}$, given phenotype and covariate information ($\boldsymbol{X}$, $\boldsymbol{Y}$), under the null hypothesis of no association and no linkage, with sample structure incorporated in either $E_0(\boldsymbol{G}|\boldsymbol{X}, \boldsymbol{Y})$ or $\mathrm{Var}_0(\boldsymbol{G}|\boldsymbol{X}, \boldsymbol{Y})$ or both[4, 6, 27–29]. For example, MASTOR[29] is a retrospective association method, applicable to structured samples, with a test statistic that is constructed using a (prospective) LMM for the phenotype. Assessment of significance for MASTOR is then based on a quasi-likelihood model for $\boldsymbol{G}$, under the null hypothesis of no association and no linkage, given by

$$\mathrm{E}_0(\boldsymbol{G}|\boldsymbol{X}, \boldsymbol{Y})=\boldsymbol{X}\boldsymbol{\alpha}, \text{ and } \mathrm{Var}_0(\boldsymbol{G}|\boldsymbol{X}, \boldsymbol{Y})=\sigma_g^2\boldsymbol{\Phi}, \quad (4)$$

where $\boldsymbol{\alpha}$; is a k-dimensional vector for the fixed effects of covariates (typically representing population structure) on the genotype, and σ_g^2 represents the variance of G_i for an outbred individual.

Theoretical results on retrospective vs. prospective analysis

Intuitively, analysis with a retrospective model is a more natural approach than is analysis with a prospective model in the case of either phenotype-based or phenotype- and covariate-based ascertainment. This can be seen in the following theoretical result: Assume that ascertainment is either phenotype-based or phenotype- and covariate-based (not genotype-based), by which we mean that conditional on ($\boldsymbol{X}$, $\boldsymbol{Y}$), ascertainment is independent of $\boldsymbol{G}$. If a model, call it $\mathcal{M}_P$, is a correctly-specified prospective model for $\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{G}$ in an unascertained sample, then $\mathcal{M}_P$ generally becomes a misspecified model in an ascertained sample. In contrast, if $\mathcal{M}_R$ is a correctly-specified retrospective model for $\boldsymbol{G}|\boldsymbol{X}$, $\boldsymbol{Y}$ in an unascertained sample, it remains a correctly-specified model in an ascertained sample.

To see why this result holds, let $\boldsymbol{S}$ be the vector of n sampling indicators, with $S_i = 1$ if individual i is included in the sample, and $S_i = 0$ otherwise. In a retrospective analysis, the model for an ascertained sample is $P(\boldsymbol{G}|\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{S} = 1)$, which turns out to be the same as the model for a population-based sample, $P(\boldsymbol{G}|\boldsymbol{X}, \boldsymbol{Y})$, due to the following relation

$$P(\boldsymbol{G}|\boldsymbol{X},\boldsymbol{Y},\boldsymbol{S}=1)=\frac{P(\boldsymbol{G}|\boldsymbol{X},\boldsymbol{Y})P(\boldsymbol{S}=1|\boldsymbol{G},\boldsymbol{X},\boldsymbol{Y})}{P(\boldsymbol{S}=1|\boldsymbol{X},\boldsymbol{Y})}=P(\boldsymbol{G}|\boldsymbol{X},\boldsymbol{Y}), \quad (5)$$

where the last equality follows from the conditional independence between ***S*** and ***G*** given ***X*** and ***Y***. In a prospective analysis, however, the population-based model $P(\boldsymbol{Y}|\boldsymbol{G},\boldsymbol{X})$ is generally not equal to $P(\boldsymbol{Y}|\boldsymbol{G},\boldsymbol{X},\boldsymbol{S}=1)$, and is therefore misspecified for ascertained data.

In general, model misspecification could be expected to lead to either inadequate control of type 1 error or loss of power or both. The above theoretical result suggests that ascertainment poses a greater risk of model misspecification for prospective analysis than for retrospective analysis. Even in the absence of ascertainment, however, model misspecification is still an important concern, with prospective analysis expected to be more sensitive to misspecification of the phenotype model while retrospective analysis would be expected to be more sensitive to misspecification of the genotype model under the null hypothesis. Arguably, the problem of correctly specifying the phenotype model is inherently harder than that of correctly specifying the genotype model, because Mendelian genetics provides strong information on a plausible null model for ***G***. In contrast, correct calibration of a prospective test could rely on accurate estimation of one or more variance component (VC) parameters (for example, heritability of the trait) in the phenotypic model, which is frequently challenging to achieve. In general, underestimated heritability is expected to inflate type 1 error because sample structure is not adequately accounted for, and over-estimated heritability is expected to lead to deflated type 1 error and hence power loss. While statistically consistent estimators of the VC parameters may be available, the sample size required for VC parameters to be accurately estimated is typically much larger than that required for comparable accuracy of the mean parameters (for example, parameters for covariate effects). For this reason, it is often not clear whether we have enough information in a given data set to reliably infer the VC parameters. In particular, binary trait data tend to contain a relatively low amount of information on unknown parameters compared to data on quantitative traits[30–32], making it especially challenging to estimate VC parameters for binary traits. In contrast, a retrospective test is able to circumvent this problem by relying on a genotypic model, so while it can make use of VC parameters for the trait model, it is less sensitive to whether or not they are well-estimated.

We have so far argued that retrospective analysis should theoretically provide greater robustness, in terms of type 1 error control and power, than prospective analysis, in the presence of phenotype model misspecification, including misspecification due to ascertainment. One can further argue that, in the presence of ascertainment, a prospective analysis can lose some information on association, where that additional information should, in principle, be available to a retrospective analysis. To see this, first recall the widely-accepted statistical principle that conditioning an analysis on an ancillary variable results in no loss of information, where a variable is defined to be ancillary if its distribution does not depend on the parameter of interest[33]. Using this principle, we first note that, for unascertained studies, both prospective and retrospective modeling are fully informative for inference on the association between the genotype and phenotype. This is because when the

sample is not ascertained, and hence is a random sample from the population, the joint distribution of (***X***, ***Y***) carries no information on whether the trait is associated with *G*. Therefore, (***X***, ***Y***) constitutes ancillary data, and the retrospective approach is fully informative. Similarly, for an unascertained sample, (***X***, ***G***) is also ancillary, and the prospective approach is also fully informative. However, when the sample is ascertained, the prospective analysis can lose information because (***X***, ***G***) may no longer be ancillary as its joint distribution may carry information on whether the phenotype is associated with *G*.

To illustrate this, we consider an extreme hypothetical example, in which the phenotype is a rare disease and sex (encoded as ***X***) is a strong covariate. Suppose the disease only occurs in males, a fact that was unknown prior to the case-control study, in which all cases are males and the control group contains 50% males and 50% females. When ***G*** represents a causal variant, one would expect to see a correlation between ***X*** and ***G*** in the data due to ascertainment, although such correlation may not exist in the population. In other words, ascertainment has introduced information about whether or not ***G*** and ***Y*** are associated into the joint distribution of (***X***, ***G***). Such information is discarded by a prospective association test, which considers ***X*** and ***G*** as fixed. A full retrospective analysis, on the other hand, is insensitive to ascertainment, because (***X***, ***Y***) remains ancillary for association with ***G***, as in an unascertained sample. Therefore, a retrospective analysis does not incur loss of information by conditioning on ***X*** and ***Y***, whether the sample is ascertained or not. In practice, however, it is not clear whether any of the previously-proposed retrospective tests are actually making use of the additional information on association contained in the conditional distribution of ***G***|***X***, ***S*** = 1. This could potentially be a topic for future work.

CARAT and CERAMIC: retrospective methods for binary trait analysis with covariates in structured samples

CARAT and CERAMIC are recently-proposed, retrospective, binary-trait, association testing methods that are designed to have the following desirable features: (1) phenotype modeled as binary, with appropriate dependence between the mean and variance; (2) appropriate correction for relevant covariates; (3) appropriate correction for sample structure of various types, including related individuals, admixture, and other types of population structure; (4) both fast and accurate computations; (5) robustness to ascertainment and other types of phenotype model misspecification; and for CERAMIC, (6) ability to leverage partially missing data to increase power. We first describe CARAT, which is designed for a sample of ostensibly unrelated individuals with population structure.

CARAT starts with a quasi-likelihood model for the phenotype, in which only the conditional mean and variance of ***Y*** given genotype and covariate information are specified. It assumes the same mean structure as logistic regression, given by

$$\mathrm{E}(Y_i|\boldsymbol{X},\boldsymbol{G})=\mu_i,\ g(\mu_i)=\log\frac{\mu_i}{1-\mu_i}=\boldsymbol{X}_i\boldsymbol{\beta}+G_i\gamma,\ i=1,\cdots,n. \quad (6)$$

The covariance structure of CARAT, which combines aspects of both a logistic regression model and a LMM, is given by

$$\mathrm{Var}(\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{G})=\boldsymbol{\Gamma}^{1/2}\sum\boldsymbol{\Gamma}^{1/2},\ \text{with}\ \sum=\xi\boldsymbol{\Phi}+(1-\xi)\boldsymbol{I}, \quad (7)$$

where $\boldsymbol{\Gamma}$ is an $n \times n$ diagonal matrix with ith diagonal element equal to $\mu_i(1 - \mu_i)$, and $0 \leq \xi \leq 1$ is an unknown VC parameter. The matrix Σ reflects the covariance structure of LMM, as given in equation 2, with the parameter ξ analogous to the heritability parameter. The matrix $\boldsymbol{\Gamma}$ allows the variance of the binary phenotype Y_i to depend on its mean in a way that aligns with the Bernoulli distribution as assumed by the logistic regression model (see equation 1). The estimate, $(\hat{\beta}_0, \hat{\xi}_0)$, of the parameter (β, ξ), under the null hypothesis that $\gamma = 0$, is obtained by solving a system of estimating equations. The CARAT association test statistic is based on the quasi-score statistic

$$U_0:=\boldsymbol{G}^T\hat{\boldsymbol{\Gamma}}_0^{1/2}\hat{\sum}_0^{-1}\hat{\boldsymbol{\Gamma}}_0^{-1/2}(\boldsymbol{Y}-\hat{\boldsymbol{\mu}}_0), \quad (8)$$

where $\hat{\boldsymbol{\mu}}_0$, $\hat{\boldsymbol{\Sigma}}_0$ and $\hat{\boldsymbol{\Lambda}}_0$ are $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$; and $\boldsymbol{\Gamma}$, respectively, evaluated at $(\gamma, \boldsymbol{\beta}, \xi) = (0, \hat{\boldsymbol{\beta}}_0, \hat{\xi}_0)$. Significance is then assessed retrospectively, with the quasi-score statistic normalized by its retrospective variance, resulting in the CARAT statistic defined as

$$\mathrm{CARAT}:=\frac{(\boldsymbol{Z}^T\boldsymbol{G})^2}{\widehat{\mathrm{Var}}_0(U_0|\,\boldsymbol{Y},\boldsymbol{X})}=\frac{(\boldsymbol{Z}^T\boldsymbol{G})^2}{\hat{\sigma}_g^2\cdot\boldsymbol{Z}^T\boldsymbol{\Phi}\boldsymbol{Z}},\ \text{with}\ \boldsymbol{Z}=\hat{\boldsymbol{\Gamma}}_0^{1/2}\hat{\sum}_0^{-1}\hat{\boldsymbol{\Gamma}}_0^{-1/2}(\boldsymbol{Y}-\hat{\mu}_0), \quad (9)$$

where $\hat{\sigma}_g^2$ is an estimate of σ_g^2 and "Var_0" represents a variance taken under the retrospective model given in equation 4.

For samples with related individuals, an additional strength of a retrospective test is its ability to leverage partially missing data to improve power[5, 27, 29, 34]. For example, phenotype and covariate data on individuals with missing genotype can contribute information to association testing when these individuals have genotyped relatives, because the latter contain information on the missing genotypes. The retrospective approach offers a natural way to incorporate such data by using the known dependence among relatives'; genotypes under the null hypothesis, while properly taking into account uncertainty. CERAMIC[5] is a recently proposed extension of CARAT to account for familial relatedness and to incorporate partially missing data. CERAMIC incorporates the phenotype and covariate data on ungeno-typed individuals who have genotyped relatives. This is done by imputing missing genotypes based on relatives' genotypes, while downweighting imputations with low information level and correcting for additional uncertainty and dependence due to imputation. As a result, CERAMIC is able to avoid the information loss

that would occur if partially missing data were dropped from the analysis. Simulations[5] show that this approach allows CERAMIC to substantially boost the power of association testing in the presence of partially missing data.

Importance of various modeling choices for binary trait mapping

We are interested in the impact of certain modeling choices, such as logistic vs. linear mean structure, dependence of the binary trait variance on the mean, and retrospective vs. prospective assessment of significance, on type 1 error control and power for case-control association mapping, There can be a complicated interplay between these features. For example, in the context of retropsective association analysis of a binary trait, recent simulation studies[5] clearly indicate higher power for logistic instead of linear mean structure when covariates play an important role, even if the true trait model is not logistic, and in those cases, the retrospective models with logistic mean structure have higher power than both retrospective and prospective LMM. However, in the context of prospective association analysis, one study found that methods with logistic mean structure and mean-variance dependence could have poor control of type 1 error, while type 1 error of LMM was well-controlled, when the true trait model is not logistic[4], while another study found that, in the presence of strong population stratification, LMM could have poor control of type 1 error while the type 1 error of logistic GLMM was well-controlled[15]. The seeming inconsistency of these latter two results might reflect the difficulty of controlling type 1 error in a prospective method with a misspecified trait model. Recent results[4], showing that CARAT is consistently more powerful than both prospective LMM and prospective logistic regression with 10 principal components included as covariates, presumably reflect the advantages of both the use of binary mean-variance structure and retrospective analysis.

Simulation study

We focus on the relative performance of prospective and retrospective methods in the presence of model misspecification. An intriguing finding from previous work[5] in the context of related individuals with known pedigree, is that when the true binary trait model contains important non-confounding covariates that are left out of the fitted model, retrospective LMM can have substantially higher power than prospective LMM. We conduct a simulation study comparing CARAT, prospective LMM, and retrospective LMM (i.e. MASTOR) in different settings of population structure and admixture, with the traits generated according to a liability threshold trait model, with major gene effects following a dominant model with epistasis, so that all 3 analysis methods (CARAT, LMM and MASTOR) are based on misspecified trait models.

To compare the empirical performance of the prospective and retrospective approaches, we simulate genotype, covariate and binary phenotype data on a sample of individuals from a structured population. The two types of structure we consider are (1) population stratification with two subpopulations and (2) two-subpopulation admixture. In each setting, we simulate 10,000 non-causal SNPs, which are used to correct for population structure. In addition, we generate two causal SNPs, which are assumed to influence the phenotype with epistasis: individuals holding at least one copy of the minor allele at both SNPs have an

elevated disease risk compared to those who do not. For our power studies, we test association at one of the two causal SNPs, treating the genotype at the other causal SNP as unobserved. For all SNPs, allele frequencies in different sub-populations are generated according to the Balding-Nichols model with fixation index $F = .01$. Conditional on genotype, covariates and ancestry, binary phenotypes are simulated according to a liability threshold model. In the population stratification model, the binary trait is generated as

$$\begin{aligned} &Y_i = 1 \text{if and only if} L_i > 0, \\ &\text{with} L_i = \boldsymbol{X}_i\boldsymbol{\beta} + \lambda \cdot 1(G_{1,i} > 0, G_{2,i} > 0) + e_i, \end{aligned} \quad (10)$$

where $Y_i = 1$ indicates that individual i is a case (affected), and $Y_i = 0$ indicates that i is a control (unaffected); $\boldsymbol{X_i}$ is an 1 × 4 covariate vector, which includes an intercept term; $\boldsymbol{\beta}$ contains the fixed covariate effects; $G_{1,i}$ and $G_{2,i}$ encode the individual's genotypes at the two causal SNPs; λ is a parameter scaling the effect of the causal SNPs; $1(G_{1,i} > 0, G_{2,i} > 0)$ is an indicator function which takes value 1 when both the causal SNPs have at least one copy of the minor allele; and $e_i \sim_{i.i.d.} N(0, \sigma_e^2)$ represents independent noise.

In the population stratification model, an ascertainment scheme is used in which a certain proportion of cases is sampled from subpopulation 1 with the remainder from subpopulation 2, where this proportion is varied from 50-80%.

In the admixture model, the binary trait is generated as

$$\begin{aligned} &Y_i = 1 \text{if and only if} L_i > 0, \\ &\text{with} L_i = \boldsymbol{X}_i\boldsymbol{\beta} + \lambda \cdot 1(G_{1,i} > 0, G_{2,i} > 0) + A_{i\rho} + e_i, \end{aligned} \quad (11)$$

where A_i is the proportion of ancestry from population 1 for the ith admixed individual, and $A_{i\rho}$ is an ancestry effect on the phenotype. In the admixture model, we vary the relative impact of ancestry versus covariates on the phenotype, where the relative impact is defined to be the variance, on the liability scale, explained by the ancestry effects, divided by the total variance explained by ancestry and the covariates. These two models are, respectively, the “Liability Threshold Trait Model with 2 Subpopulations” and the “Liability Threshold Trait Model with Admixture” in a previously published work[4], and further details of the simulation models can be found there.

For each simulation scenario, we investigate the performance of three methods: CARAT, (prospective) LMM, and MASTOR (i.e., retrospective LMM). Note that software for both CARAT and MASTOR is freely-downloadable from http://galton.uchicago.edu/~mcpeek/software/index.html. In our simulations, all three methods correctly control type 1 error (results not shown). To compare power of retrospective and prospective analysis when the trait model is misspecified, it is particularly instructive to compare MASTOR and LMM, which are based on the same LMM score function and differ only in having a retrospective (MASTOR) or prospective (LMM) variance calculation. As shown in Tables 1 and 2,

MASTOR has a substantial power advantage over LMM when ascertainment or population admixture has a strong effect on the phenotype, while remaining as powerful as LMM when such effects are relatively weak. Thus, our simulation results indicate that the retrospective approach can lead to improved power compared to the prospective approach.

Compared with MASTOR, CARAT assumes a more appropriate model for binary traits by incorporating covariate and genetic effects on the logit scale, and by modeling mean-variance dependency. Tables 1 and 2 show the substantial power advantage of CARAT over MASTOR, particularly when covariates have a strong effect on the trait.

Discussion

For association mapping of a binary trait, use of a LMM designed for a quantitative trait, while computationally convenient, is generally not statistically efficient. When covariates play an important role in the association analysis of a binary trait, the use of a trait model having mean and variance structure appropriate to a binary trait, instead of a linear model, can improve power. However, in the context of prospective association analysis, such an approach can have poor type 1 error control when the trait model is misspecified[4]. A retrospective approach is able to circumvent this problem by relying on a genotypic model, which is arguably easier to specify correctly than a phenotypic model. Compared to prospective tests, retrospective tests have the advantage of being more robust to trait model misspecification, in terms of both type 1 error control and power. Thus, in particular, retrospective analysis can overcome some of the limitations of the additive polygenic model by providing a degree of robustness to the failure of the assumptions of that model.

We give a theoretical justification for retrospective analysis being the more natural approach in the case of ascertained data, which is a very common situation in case-control association analysis. We argue that if a prospective model is correctly specified for a population-based sample, then, in general, it becomes a misspecified model under phenotype-based ascertainment, while if a retrospective model is correctly specified for a population-based sample, then it does not change under phenotype and/or covariate-based ascertainment. We provide further simulation results showing that retrospective association analysis tends to be more robust to model misspecification, in terms of power, than prospective association analysis. Finally, an additional advantage to retrospective association analysis is that it provides a natural way to incorporate partially missing information into the analysis, while properly taking into account uncertainty.

We describe recently-proposed retrospective binary-trait association mapping methods, CARAT and CERAMIC, that are developed to achieve the following goals: (1) phenotype modeled as binary, with appropriate dependence between the mean and variance; (2) appropriate correction for relevant covariates; (3) appropriate correction for sample structure of various types, including related individuals, admixture, and other types of population structure; (4) both fast and accurate computations; (5) robustness to ascertainment and other types of phenotype model misspecification; and for CERAMIC, (6) ability to leverage partially missing data to increase power. The retrospective approach allows these methods to retain type 1 error control across a wide variety of settings, including misspecified link

function (e.g., probit vs. logit), misspecified model for genetic effects on the trait (e.g. dominant effects with epistasis instead of effects that are additive both within and between loci), misspecified variance structure (e.g., failure to model dominance variance), missing covariates, various types of population structure, and ascertainment. In simulations, we show that CARAT tends to have power that is as high or higher than that of the retrospective LMM method MASTOR, which itself tends to have power that is as high or higher than that of the prospective LMM.

Acknowledgments

This study was supported by National Institutes of Health (NIH) grant R01 HG001645 (to M.S.M.).

References

1. Kang HM, Sul JH, Service SK, Zaitlen NA, Kong Sy, Freimer NB, Sabatti C, Eskin E, et al. Variance component model to account for sample structure in genome-wide association studies. Nature Genetics. 2010; 42:348–354. [PubMed: 20208533]
2. Consortium IMSG, 2 WTCCC. Genetic risk and a primary role for cell-mediated immune mechanisms in multiple sclerosis. Nature. 2011; 476:214–219. [PubMed: 21833088]
3. Yang J, Zaitlen NA, Goddard ME, Visscher PM, Price AL. Advantages and pitfalls in the application of mixed-model association methods. Nature Genetics. 2014; 46:100–106. [PubMed: 24473328]
4. Jiang D, Zhong S, McPeek MS. Retrospective Binary-Trait Association Test Elucidates Genetic Architecture of Crohn Disease. The American Journal of Human Genetics. 2016; 98:243–255. [PubMed: 26833331]
5. Zhong S, Jiang D, McPeek MS. CERAMIC: case-control association testing in samples with related individuals based on retrospective mixed model analysis with adjustment for covariates. Under review. 2016
6. Bourgain C, Hoffjan S, Nicolae R, Newman D, Steiner L, Walker K, Reynolds R, Ober C, McPeek MS. Novel case-control test in a founder population identifies P-selectin as an atopy-susceptibility locus. The American Journal of Human Genetics. 2003; 73:612–626. [PubMed: 12929084]
7. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. Nature Genetics. 2006; 38:904–909. [PubMed: 16862161]
8. Nerlove M, Press JS. Univariate and Multivariate Log-linear and Logistic Models. Rand Corporation. 1973
9. Schall R. Estimation in generalized linear models with random effects. Biometrika. 1991; 78:719–727.
10. Breslow NE, Clayton DG. Approximate inference in generalized linear mixed models. Journal of the American statistical Association. 1993; 88:9–25.
11. Wolfinger R, O'connell M. Generalized linear mixed models a pseudo-likelihood approach. Journal of statistical Computation and Simulation. 1993; 48:233–243.
12. Rodriguez G, Goldman N. Improved estimation procedures for multilevel models with binary response: a case-study. Journal of the Royal Statistical Society: Series A (Statistics in Society). 2001; 164:339–355.
13. Liu D, Ghosh D, Lin X. Estimation and testing for the effect of a genetic pathway on a disease outcome using logistic kernel machine regression via logistic mixed models. BMC Bioinformatics. 2008; 9:292. [PubMed: 18577223]
14. Bolker BM, Brooks ME, Clark CJ, Geange SW, Poulsen JR, Stevens MHH, White JSS. Generalized linear mixed models: a practical guide for ecology and evolution. Trends in Ecology & Evolution. 2009; 24:127–135. [PubMed: 19185386]
15. Chen H, Wang C, Conomos MP, Stilp AM, Li Z, Sofer T, Szpiro AA, Chen W, Brehm JM, Celedón JC, et al. Control for population structure and relatedness for binary traits in genetic association

studies via logistic mixed models. The American Journal of Human Genetics. 2016; 98:653–666. [PubMed: 27018471]

16. Raudenbush SW, Yang ML, Yosef M. Maximum likelihood for generalized linear models with nested random effects via high-order, multivariate Laplace approximation. Journal of computational and Graphical Statistics. 2000; 9:141–157.
17. Rabe-Hesketh S, Skrondal A, Pickles A. Reliable estimation of generalized linear mixed models using adaptive quadrature. The Stata Journal. 2002; 2:1–21.
18. Pinheiro JC, Chao EC. Efficient Laplacian and adaptive Gaussian quadrature algorithms for multilevel generalized linear mixed models. Journal of Computational and Graphical Statistics. 2006:58–81.
19. Gilks WR, Richardson S, Spiegelhalter DJ. Introducing Markov chain Monte Carlo. Markov Chain Monte Carlo in Practice. 1996; 1:19.
20. Clayton, DG. Markov Chain Monte Carlo in Practice. Springer; 1996. Generalized linear mixed models; p. 275-301.
21. Berger J. The case for objective Bayesian analysis. Bayesian Analysis. 2006; 1:385–402.
22. Cowles MK, Carlin BP. Markov chain Monte Carlo convergence diagnostics: a comparative review. Journal of the American Statistical Association. 1996; 91:883–904.
23. Lin DY, Tao R, Kalsbeek WD, Zeng D, Gonzalez F, Fernández-Rhodes L, Graff M, Koch GG, North KE, Heiss G. Genetic association analysis under complex survey sampling: the Hispanic Community Health Study/Study of Latinos. The American Journal of Human Genetics. 2014; 95:675–688. [PubMed: 25480034]
24. Hayeck TJ, Zaitlen NA, Loh PR, Vilhjalmsson B, Pollack S, Gusev A, Yang J, Chen GB, Goddard ME, Visscher PM, et al. Mixed model with correction for case-control ascertainment increases association power. The American Journal of Human Genetics. 2015; 96:720–730. [PubMed: 25892111]
25. Weissbrod O, Lippert C, Geiger D, Heckerman D. Accurate liability estimation improves power in ascertained case-control studies. Nature Methods. 2015; 12:332–334. [PubMed: 25664543]
26. Stanhope SA, Abney M. GLOGS: a fast and powerful method for GWAS of binary traits with risk covariates in related populations. Bioinformatics. 2012; 28:1553–1554. [PubMed: 22522135]
27. Thornton T, McPeek MS. Case-control association testing with related individuals: a more powerful quasi-likelihood score test. The American Journal of Human Genetics. 2007; 81:321–337. [PubMed: 17668381]
28. Thornton T, McPeek MS. ROADTRIPS: case-control association testing with partially or completely unknown population and pedigree structure. The American Journal of Human Genetics. 2010; 86:172–184. [PubMed: 20137780]
29. Jakobsdottir J, McPeek MS. MASTOR: mixed-model association mapping of quantitative traits in samples with related individuals. The American Journal of Human Genetics. 2013; 92:652–666. [PubMed: 23643379]
30. Xu S, Atchley WR. Mapping quantitative trait loci for complex binary diseases using line crosses. Genetics. 1996; 143:1417–1424. [PubMed: 8807312]
31. Duggirala R, Williams JT, Williams-Blangero S, Blangero J. A variance component approach to dichotomous trait linkage analysis using a threshold model. Genetic Epidemiology. 1997; 14:987–992. [PubMed: 9433612]
32. Wijsman EM, Amos CI. Genetic analysis of simulated oligogenic traits in nuclear and extended pedigrees: summary of GAW10 contributions. Genetic Epidemiology. 1997; 14:719–735. [PubMed: 9433569]
33. Birnbaum A. On the foundations of statistical inference. Journal of the American Statistical Association. 1962; 57:269–306.
34. McPeek MS. BLUP genotype imputation for case-control association testing with related individuals and missing data. Journal of Computational Biology. 2012; 19:756–765. [PubMed: 22697245]

Table 1
Empirical Power of LMM, MASTOR and CARAT with Population Stratification

Proportion of Cases From[a]		Empirical Power of		
Population 1	**Population 2**	**LMM**	**MASTOR**	**CARAT**
50%	50%	.57	.57	.70
60%	40%	.55	.56	.65
70%	30%	.53	.55	.60
80%	20%	.44	.48	.50

Empirical power is based on 5,000 replicates, so an upper bound for the standard error is .007 for every entry in the table.

[a] A sample of 2,000 individuals, with 1,000 from each of the two subpopulations and with equal numbers of cases and controls, is ascertained based on the specified proportions of cases from each subpopulation.

Table 2
Empirical Power of LMM, MASTOR and CARAT with Population Admixture

% Variance due to[a]		Empirical Power of		
Ancestry	**Covariates**	**LMM**	**MASTOR**	**CARAT**
0%	100%	.70	.70	.79
20%	80%	.59	.64	.64
40%	60%	.64	.69	.69
60%	40%	.59	.66	.69
80%	20%	.62	.71	.74
100%	0%	.58	.67	.70

Empirical power is based on 5,000 replicates, so an upper bound for the standard error is .007 for every entry in the table.

[a]The percentages are defined to be the variance, on the liability scale, explained by either the ancestry effects or the covariate effects, divided by the total variance explained by the two types of effects, indicating the relative impact of ancestry versus covariates on the phenotype.