On the validity of within-nuclear-family genetic association analysis in samples of extended families

Alexandre Bureau^{1,2} and Thierry Duchesne³

Alexandre Bureau: alexandre.bureau@msp.ulaval.ca

¹Département de médecine sociale et préventive, Pavillon Vandry, room 2457, 1050 rue de la Médecine Université Laval, Québec, QC, G1V 0A6, Canada

²Centre de recherche de l'Institut Universitaire en Santé Mentale de Québec, Beauport (QC), G1J 2G3, Canada

³Département de mathématiques et de statistique, Pavillon Alexandre-Vachon, room 1056, 1045 av. de la Médecine Université Laval, Québec, QC, G1V 0A6, Canada

Abstract

Splitting extended families into their component nuclear families to apply a genetic association method designed for nuclear families is a widespread practice in familial genetic studies. Dependence among genotypes and phenotypes of nuclear families from the same extended family arises because of genetic linkage of the tested marker with a risk variant or because of familial specificity of genetic effects due to gene-environment interaction. This raises concerns about the validity of inference conducted under the assumption of independence of the nuclear families. We indeed prove theoretically that, in a conditional logistic regression analysis applicable to disease cases and their genotyped parents, the naive model-based estimator of the variance of the coefficient estimates underestimates the true variance. However, simulations with realistic effect sizes of risk variants and variation of this effect from family to family reveal that the underestimation is negligible. The simulations also show the greater efficiency of the model-based variance estimator compared to a robust empirical estimator. Our recommendation is therefore to use the model-based estimator of variance for inference on effects of genetic variants.

Keywords

Conditional logistic regression; Empirical variance estimation; Genetic linkage; Pseudo-controls; Wald test

1 Introduction

Methods to estimate association parameters and test the null hypothesis of absence of association between dichotomous phenotypes and genetic markers in samples of nuclear families are well developed (see for instance Ziegler and König, 2010, chap. 12). Within-family analysis, conditional on phenotype and genotype observed in each family, is well

known to protect against confounding due to population stratification. Family samples recruited for genetic studies often include extended families containing multiple nuclear families spanning three or more generations. For instance, the Eastern Quebec Schizophrenia and Bipolar Disorder Study is conducted on a sample of 48 kindreds containing 202 nuclear families (Bureau, Chagnon, Croteau, Fournier, Roy, Paccalet, Merette, and Maziade, 2013). Existing methods to analyze association in extended families provide only hypothesis tests, not effect size estimates (Martin, Bass, and Kaplan, 2001, Lange, DeMeo, Silverman, Weiss, and Laird, 2004, Chen, Manichaikul, and Rich, 2009, Bureau, Croteau, Chagnon, Roy, and Maziade, 2014). A widespread alternative to these methods tailored for extended families is to split extended families into nuclear families and apply one of the wide array of available methods. Many software packages such as FBAT (standing for Family-Based Association Test, www.biostat.harvard.edu/fbat/default.html), the TDTEX program of SAGE (darwin.cwru.edu/sage/) and Unphased (Dudbridge, 2008) accept pedigree files containing extended pedigrees, split them automatically into nuclear families and run nuclear family association analyses on the resulting nuclear families. Inference on the effect size of a variant via estimation of regression coefficients along with Wald tests and confidence intervals has been developed only for nuclear families, forcing this splitting of extended families to perform such inference. Genotypes and phenotypes of nuclear families from the same extended family however exhibit dependence from various causes, and the extent to which this dependence may compromise the validity of the within-nuclear-family analyses has not been carefully studied. FBAT uses an empirical estimator of the variance of a score statistic, robust to dependence among nuclear families from the same extended family, but SAGE's TDTEX program and Unphased (which implements Wald inference on regression models) analyze the nuclear families as if they were independent. We examine in this article the impact of dependence from two sources: genetic linkage to a nearby risk locus and interaction with family-specific environment.

When testing the null hypothesis of no association between a marker and the trait in presence of genetic linkage to a nearby risk locus, it has long been recognized that correlation between the transmitted alleles of siblings inflates the Type I error of association test statistics treating the siblings as independent within nuclear families (Martin, Kaplan, and Weir, 1997). This problem has been solved in nuclear family analysis by using a robust estimator of variance of the score statistic (Lake, Blacker, and Laird, 2000) or the coefficient estimate of a logistic model (Siegmund, Langholz, Kraft, and Thomas, 2000), by estimating identity-by-descent probabilities (Martin, Bass, Hauser, and Kaplan, 2003) or by further conditioning on equivalence classes of inheritance vectors under linkage in the logistic conditional likelihood (Dudbridge, 2008, Dudbridge, Holmans, and Wilson, 2011). It has been recognized that genetic linkage also induces dependence among the genotypes of subjects in different nuclear families from the same extended family, and that this could bias simple averaging of nuclear family-level statistics across nuclear families (Martin et al., 2001). Valid score tests of association in the presence of linkage have been obtained by modifying the weighting of nuclear-family contributions (Martin et al., 2001), simulation of the null distribution (Allen-Brady, Wong, and Camp, 2006) or use of identity-by-descent proportions estimated from surrounding markers in the estimation of the variance of the score statistic (Chen et al., 2009). The impact on variance estimation of coefficient estimates

and Wald tests of applying within-nuclear-family logistic conditional likelihood inference to nuclear families from extended families in the presence of linkage has however never been investigated to our knowledge.

If one is instead interested in testing the null hypothesis of no association and no linkage, dependence among relatives due to linkage is no longer a source of Type I error inflation, but instead information contributing to power (Siegmund et al., 2000). But even then, the phenotypes of relatives from different nuclear families from the same extended family could further depend on each other because of a common environment modifying the effect of susceptibility alleles (gene-environment interaction)(Lynch and Walsh, 1998, Chapter 7). The impact on within-nuclear-family association analysis of this second source of dependence among nuclear families has not been studied previously.

When Wald-type inference is valid, the splitting of extended families into nuclear families required to perform such inference may lead to lower power compared to a score test designed for extended families. The Generalized Disequilibrium Test (GDT), a score statistic for extended families derived from a conditional likelihood, has been found more powerful than score tests for nuclear families such as the FBAT applied on the same samples of extended families under a range of simulation scenarios (Chen et al., 2009). To our knowledge, a similar comparison with a Wald test has not been performed.

Our objectives are 1) to investigate the direction and magnitude of the bias in the estimation of the variance of logistic regression coefficient estimates when applying a within-nuclear family analysis to a sample of nuclear families obtained by splitting extended families within which dependence is present, 2) evaluate whether an empirical estimator of variance robust to dependence among nuclear families from the same extended family should be preferred to an estimator obtained by naively assuming independence among nuclear families and 3) to compare the power of the Wald test on nuclear families with an extended families score test.

With genotyped parents and a single affected subject per nuclear family (case-parent trios), inference can be carried out under a logistic conditional likelihood where three pseudo-controls are constructed from the parental untransmitted genotypes (Self, Longton, Kopecky, and Liang, 1991). Under this model, we represent dependence among nuclear families due to either genetic linkage or environment-driven extended family-specific effect by random effects. This allows us to prove theoretically that the model-based variance of the coefficient estimate of the logistic model underestimates the true variance. The extent of this underestimation is investigated by simulation. In simulations, we also consider the case of multiple affected subjects per nuclear family, where dependence among siblings in presence of linkage to a nearby risk locus is dealt with using the conditioning on inheritance vector of Dudbridge (2008), but dependence across nuclear families is ignored. We also illustrate the behaviour of the variance estimates on genetic markers genotyped in the Eastern Quebec Schizophrenia and Bipolar Disorder Study.

2 Methods

2.1 Within-nuclear-family conditional inference given parental genotype

Consider a study with K extended families, where each extended family c contains n_c nuclear families with at least one affected child (there may also be nuclear families without any affected children, but these don't contribute to the conditional likelihood). Let us first assume exactly one affected child per nuclear family and let the genotype M of that affected child at a genetic marker take a value m belonging to the set S_{cs} of possible genotypes transmitted by his parents in nuclear family s from extended family c . We denote $X^{(m)}$ the number of copies of the minor (least frequent) allele in marker genotype $M = m$ and use X_{cs} as a condensed notation for the allele count of the marker genotype actually observed in the affected subject of nuclear family s from extended family c . Assuming a logistic regression model holds, the disease probability is then given by

$$\Pr(Y_{cs}=1|X_{cs}) = \frac{\exp\{\alpha + \beta X_{cs}\}}{1 + \exp\{\alpha + \beta X_{cs}\}}, \quad (1)$$

assuming the intercept α and the regression coefficient β between X and the disease status Y is constant in all nuclear families across all extended families. The conditional likelihood of the affected children genotypes given the parental genotypes in their respective nuclear families takes the form

$$\prod_{c=1}^K \prod_{s=1}^{n_c} \frac{\exp(\beta X_{cs})}{\sum_{\ell \in S_{cs}} \exp(\beta X_{cs}^{(\ell)})} \quad (2)$$

This likelihood is the same as the conditional likelihood of logistic regression under a matched case-control design, with the genotypes not transmitted to the child playing the role of pseudo-controls. Indeed, the case-parent trio design is also called the case-pseudosib design (Witte, Gauderman, and Thomas, 1999). The results on the variance of the coefficient estimates which we derive under this model therefore apply to association analysis conditional on parental genotypes, as long as there is a single affected child per nuclear family. We consider arbitrary numbers of affected children per nuclear families in Section 2.4. The assumption of a rare disease is needed to satisfy the implicit assumption that pseudo-sibs receiving untransmitted genotypes would be unaffected (see Appendix 1 of Witte et al. (1999) for an analysis of this issue). From this likelihood we estimate β by solving the score equation

$$U(\beta) = \frac{1}{K} \sum_c \sum_s \left\{ X_{cs} - \sum_{h \in S_{cs}} X_{cs}^{(h)} \frac{\exp(\beta X_{cs}^{(h)})}{\sum_{\ell \in S_{cs}} \exp(\beta X_{cs}^{(\ell)})} \right\} = 0. \quad (3)$$

If we let $\hat{\beta}$ denote the solution to (3), Wald-type inferences on β can be conducted if a consistent estimator of the variance of $\hat{\beta}$ is available. Let $I(\beta)$ denote the observed information:

$$I(\beta) = -\frac{\partial}{\partial \beta} U(\beta) = \frac{1}{K} \sum_c \sum_s \sum_{h \in S_{cs}} w_{cs}^{(h)} (X_{cs}^{(h)} - \bar{X}_{cs})^2, \quad (4)$$

where

$$w_{cs}^{(h)} = \frac{\exp(\beta X_{cs}^{(h)})}{\sum_{\ell \in S_{cs}} \exp(\beta X_{cs}^{(\ell)})} \quad (5)$$

and $\bar{X}_{cs} = \sum_{h \in S_{cs}} w_{cs}^{(h)} X_{cs}^{(h)}$. We then define the model-based variance estimator of $\hat{\beta}$ as $\text{var}_{\text{MB}}(\hat{\beta}) = K^{-1} I^{-1}(\hat{\beta})$.

We note that the likelihood (2) continues to apply when the intercept α takes a different value in each nuclear family, as the α s factor out of the likelihood (Neuhaus and Jewell, 1990).

2.2 Dependence among nuclear families from the same extended family

Likelihood (2) does not apply when β varies between extended and/or nuclear families. With extended families, the assumption of a constant β does not hold if A) the tested genetic marker M is not itself influencing the risk of disease but is instead in genetic linkage with a nearby risk locus G or B) the effect of the risk variant depends on the family-specific environment. We consider the consequences of these two phenomena separately.

2.2.1 Dependence due to genetic linkage—When the genetic variant being tested is a marker with no effect on disease risk but is linked to a locus influencing the risk of disease through model (1), the disease risk becomes

$$\Pr(Y_{cs}=1|X_{cs}) = \frac{\exp\{\alpha\}}{1+\exp\{\alpha\}} \pi_0(X_{cs}) + \frac{\exp\{\alpha+\beta\}}{1+\exp\{\alpha+\beta\}} \pi_1(X_{cs}) + \frac{\exp\{\alpha+2\beta\}}{1+\exp\{\alpha+2\beta\}} \pi_2(X_{cs}) \quad (6)$$

where $\pi_v(X) = P[V=v|X]$, the probability of v copies of the minor allele in the risk locus genotype G given X in the marker locus genotype M . This is not in general a logistic model. However, a logistic model can be a good approximation. In particular, if the marker minor allele frequency is sufficiently low that the frequency of $X=2$ is negligible, we have a saturated mean model for the two levels $X=0, 1$ which can be written

$$\Pr(Y_{cs}=1|X_{cs}, \gamma_{cs}) = \frac{\exp\{\alpha + \gamma_{cs}^{(0)} + (\beta + \gamma_{cs}^{(1)})X_{cs}\}}{1 + \exp\{\alpha + \gamma_{cs}^{(0)} + (\beta + \gamma_{cs}^{(1)})X_{cs}\}}, \quad (7)$$

with $\gamma_{cs}^\top = (\gamma_{cs}^{(0)}, \gamma_{cs}^{(1)})$.

The terms γ_{cs} capture the deviation of the nuclear family specific regression coefficients from the population coefficients. This deviation is correlated among nuclear families in the same extended family because the alleles in genotype G at the risk locus and in genotype M at the marker locus are transmitted together on the same haplotype due to their tight linkage. The γ_{cs} can be seen as random coefficients under a logistic mixed model, with a joint distribution within extended families reflecting the dependence in genetic inheritance. We compute the impact of the dependence among γ_{cs} , $s = 1, \dots, n_c$ on the variance of the coefficient estimates for first cousin pairs in Section C of the Appendix when the marker allele frequency is low. In our simulation study, we investigate applying model (7) with higher marker allele frequencies such that the values 0, 1 and 2 of X are all frequent.

2.2.2 Dependence due to interaction with family-specific environment—Even when the tested variant is influencing directly the risk of disease, the magnitude of this effect may depend on the environment in which individuals live. Related individuals tend to live in more similar environments than individuals with whom they are not related, because of geographical proximity and familial habits. While the environment is shared to a greater extent within nuclear families, some degree of environmental similarity is likely to subsist among more distant relatives. Variation in risk variant effect β between nuclear and extended families due to interaction with family-specific environment is commonly represented by random coefficients under a logistic mixed model which can be expressed by (7). The only difference with the dependence due to genetic linkage is the interpretation of the random effects.

Obviously, the two types of random effects could be present simultaneously when testing a marker linked to a risk locus whose effect depends on the environment. We study the two effects separately to distinguish their impacts on the analysis, but under the same random effect framework.

2.3 Impact of familial dependence on inference

In presence of random coefficients, the conditional likelihood given by (2) is not valid anymore. Nevertheless, we can still estimate a regression coefficient by solving the score equation (3), which will yield, when the disease is rare, a consistent estimator $\hat{\beta}^*$ of the marginal log-odds ratio β^* defined as

$$\beta^* = \log \left\{ \frac{\Pr(Y_{cs}=1|X_{cs}=x+1)/\Pr(Y_{cs}=0|X_{cs}=x+1)}{\Pr(Y_{cs}=1|X_{cs}=x)/\Pr(Y_{csi}=1|X_{cs}=x)} \right\}, \quad (8)$$

with

$$\Pr(Y_{cs}=y|X_{cs}=x) = \int \frac{\exp\{y[\alpha + \gamma_{cs}^{(0)} + (\beta + \gamma_{cs}^{(1)})x]\}}{1 + \exp\{\alpha + \gamma_{cs}^{(0)} + (\beta + \gamma_{cs}^{(1)})x\}} dF(\gamma_{cs}).$$

Indeed, in this case we have that $E_{\beta^*}\{U(\beta^*)|\Sigma_{h \in S_{cs}} Y_h = 1 \forall c, s\} = 0$ where Y_h denotes an indicator for whether genotype h was transmitted or not. If we consider Y_h as a random variable, then it can be interpreted as the pseudo-phenotype for genotype h , which is affected (1) when h represents the actual child genotype and unaffected (0) when h represents an untransmitted genotype.

In the Appendix we show that the naive model-based variance based on the inverse of the information (4) evaluated at $\beta = \hat{\beta}^*$ is not a consistent estimator of the true asymptotic variance of $\hat{\beta}^*$. As a matter of fact, we show that

$$\text{var}_{\beta^*}(\hat{\beta}^*) = \text{var}_{\text{MB}}(\hat{\beta}^*)\{1 + K \Delta \text{var}_{\text{MB}}(\hat{\beta}^*)\} \quad (9)$$

$$= \text{var}_{\text{MB}}(\hat{\beta}^*)\{1 + \Delta I^{-1}(\hat{\beta}^*)\} \quad (10)$$

for some real number Δ that we describe in further detail in the Appendix. From (10), we see that the inflation in variance does not depend on the number of extended families K . In the special cases where X is binary or may only take on three equidistant values such as an allele count 0, 1 or 2, we show in the Appendix that $\Delta = 0$, and hence that the model-based variance underestimates the true variance in those cases. The extent of this underestimation will depend on the value of Δ and we investigate the impact of this result with samples of practical sizes in the simulation study.

Craiu, Duchesne, and Fortin (2008) advocated the use of a robust sandwich variance estimator instead of $\text{var}_{\text{MB}}(\hat{\beta}^*)$ in Wald-type tests of hypotheses about β in studies with clustered strata. In our notation, this robust sandwich variance is given by

$$\text{var}_{\text{R}}(\hat{\beta}^*) = \frac{1}{K} I^{-2}(\hat{\beta}^*) \frac{1}{K} \sum_c \left\{ \sum_s (X_{cs} - \bar{X}_{cs}) \right\}^2, \quad (11)$$

where the weights $w_{cs}^{(h)}$ to compute $\bar{X}_{cs}$ are given by (5), but with β replaced by $\hat{\beta}^*$. We also compare inferences based on $\text{var}_{\text{R}}(\hat{\beta}^*)$ and on $\text{var}_{\text{MB}}(\hat{\beta}^*)$ in the simulation study.

2.4 Arbitrary number of affected children per nuclear family

In reality, nuclear families may contain more than one affected child. The likelihood (2) has been extended to an arbitrary number A_{cs} of affected children per nuclear family as follows:

$$\prod_{c=1}^K \prod_{s=1}^{n_c} \frac{\exp \left(\sum_a^{A_{cs}} \beta X_{csa}^{(m)} \right)}{\sum_{\ell \in S_{cs}} \exp \left(\sum_a^{A_{cs}} \beta X_{csa}^{(\ell)} \right)} \quad (12)$$

where m now denotes the vector of genotypes transmitted by the parents to their affected offsprings (Dudbridge, 2008). The genotypes of affected siblings may be dependent due to genetic linkage, and this has been addressed by redefining the set S_{CS} to further condition on equivalence classes of inheritance vectors (Dudbridge, 2008). The summation is generally not over all possible permutations of the genotype vector M , and this likelihood is not equivalent to the usual conditional likelihood for logistic regression. In our simulation study, we let the number of affected siblings per nuclear family be 1 and apply likelihood (12) to assess empirically whether our results on the variance of the coefficient estimates extend to that more general setting.

2.5 Simulation study

We simulated a biallelic variant where a risk allele had population frequency of either 0.1, 0.3 or 0.5. Two other biallelic variants with an allele at the same frequency as the risk allele were simulated as markers, one strongly associated in the population with the risk variant ($r^2 = 0.8$) and the other unassociated to the risk variant ($r^2 = 0$). In our first scenario to assess the effect of dependence due to genetic linkage alone, the risk allele multiplied the odds of disease by the fixed value of 1.5 for each copy of the variant present ($\beta = \log(1.5) = 0.41$), and the intercept α in model (1) was set to achieve an overall population prevalence of 1.0% in all extended families. In a second scenario, the risk allele effect and intercept were specific to each extended family, to reflect a shared environment. This was implemented by adding zero-mean normal random effects to the intercept α and the coefficient β , independent of each other and both with standard deviation (SD) of 0.2. In this way, the extended family-specific coefficient fell between 0 and 0.81 (odds ratio = 2.25) in over 95% of families. In the remaining 5% of families, values of family-specific coefficient below 0 were roofed to 0 to avoid a flip of the risk allele, and values above 0.81 were floored to 0.81 to preserve the mean effect.

We simulated extended families from a 5-generation pedigree structure with a single founder couple in the top generation and containing 80 nuclear families with 4 children each in the bottom 3 generations. Pedigree founder genotypes were simulated assuming Hardy-Weinberg equilibrium. The haplotypes formed by the alleles at the three simulated variants were then transmitted from the founders down the pedigree to their descendants as one locus, without recombination, following Mendel's laws of inheritance. The disease status of the pedigree members in the bottom 3 generations were simulated using the penetrance corresponding to their genotype at the risk variant.

Extended families were formed by keeping from the simulated 5-generation pedigrees the nuclear families with at least one affected subject, and other nuclear families linking these affected subjects in the pedigree, mimicking the ascertainment process in familial studies where nuclear families containing affected relatives of an index case are recruited, but not nuclear families containing only unaffected relatives. We applied the ascertainment criterion of at least 5 nuclear families with at least one affected subject to include an extended family into a family sample. Simulation of pedigrees continued until a sample of 60 extended families was formed. We created variables X representing the number of copies in each subject's genotype of the risk allele for the risk variant and correlated variant, and of an

allele of same frequency for the unassociated variant. We applied the GDT using the default estimator of the variance of the score based on the kinship coefficients to the X for the risk variant in the samples of extended families and estimated the log-odds ratio β or β^* for each X under a logistic regression model of disease status by maximizing the conditional likelihood (12) on all nuclear families in a sample as if they were independent. The model-based and robust estimates of the variance of $\hat{\beta}^*$ were obtained using equations (4) and (11), respectively. We used our own implementation in R, which returns identical estimates and model-based standard errors as the Unphased package (Dudbridge, 2008). This process was repeated 5000 times.

3 Results

3.1 Simulation with dependence due to genetic linkage

In this simulation to assess the effect of within-family dependence due to genetic linkage, there were on average 5.8 affected subjects per extended family (349 affected subjects per sample of 60 extended families), with 50% of families having more than the minimum of 5 in 5 different nuclear families required for ascertainment of the family. Estimates of the coefficient β were approximately unbiased (results not shown). Results on the behaviour of the variance estimators in the simulation study are reported in Table 1. The mean of the model-based and robust variance estimates were estimated with high precision and were close to each other. We predicted the true variance based on equation (9) from the calculations outlined in Section C of the Appendix for a marker allele with frequency 0.1 with the model-based variance estimated by the mean of the model-based variance estimate. This predicted variance reported in Table 1 was only slightly inflated and fell within the 95% confidence interval for the true variance constructed from the coefficient estimates. We could not establish whether the true variance was statistically significantly underestimated by the model-based or robust variance estimators for markers strongly associated with a tightly linked risk variant ($r^2 = 0.8$) or unassociated to it ($r^2 = 0$) due to the greater variability of the sample variance (taken as unbiased estimate of the true variance) compared to the mean of the model-based variance estimate. It is however clear that any underestimation would be practically insignificant. No bias in the variance estimates was expected when estimating the coefficient of the risk variant itself, and indeed no significant evidence of bias of the variance estimates was detected.

We also noticed that the SD of the robust variance estimate was much larger than the SD of the model-based estimate (from 2 times larger at an allele frequency of 0.1 to almost 4 times larger at an allele frequency of 0.5 in Table 1). This has consequences on inference: the Type I error of the Wald test is slightly larger for a variant linked but not associated (Table 2, Allelic odds ratio = 1), although it remains close to the nominal level. Confidence interval coverage for the risk variant effect β does not however depart significantly from the nominal level with the robust variance estimate (Table 3, Fixed effect). Both confidence interval coverage and Type I error were close to nominal level with the model-based estimator. The Wald tests had slightly lower power to detect linkage and association to the risk variant than the GDT (Table 2, Allelic odds ratio: fixed at 1.5). Whether the model-based or robust variance estimator was used had no noticeable impact on power.

3.2 Simulation with dependence due to shared environment

There were on average 6.8 affected subjects per extended family (409 affected subjects per sample of 60 extended families), with 69% of families having more than the minimum of 5 in 5 different nuclear families required for ascertainment of the family. Coefficient estimates were centered around population values β^* larger than the mean of extended family-specific effects β (results not shown). In this setting with normal random effects, we were unable to compute the expected true variance based on equation (9) because this requires the distribution of the random effect conditional on the observed values of X for all family members and on the ascertainment of one case per nuclear family (see equation (17) of the Appendix), which is difficult to obtain. As in the simulation with dependence due to genetic linkage, no practically significant underestimation of the sample variance by the model-based variance was detected, the mean of the model-based and robust variance estimates were close to each other, and the SD of the robust variance estimate was 2 to 3 times larger than the SD of the model-based estimate. The confidence interval coverage for the marginal risk variant effect β^* was slightly under the nominal level for a variant with frequency 0.5 but remained close to nominal level at lower variant frequencies with either variance estimator (Table 3, Family-specific effect). Power of the Wald test and GDT was somewhat higher under this scenario than with a risk variant effect β fixed over all families, but otherwise the previous observations of slightly lower power of the Wald test compared to the GDT and no influence of the variance estimator still apply here (Table 2, Allelic odds ratio: family-specific with mean of 1.5).

3.3 Illustration on the Eastern Quebec Schizophrenia and Bipolar Disorder Study

We performed a within-nuclear-family conditional association analysis given parental genotypes on the q arm of chromosome 13, a region studied by Bureau et al. (2013) because of its linkage to a “common locus” phenotype comprising schizophrenia, bipolar disorder and schizoaffective disorder. We analyzed association between the broad version of the “common locus” phenotype and 48 single nucleotide polymorphism markers selected for their low correlation to each other in the study sample. We restricted the study sample to the nuclear families where both parents were genotyped, leaving between 66 and 90 nuclear families (mean = 81) from between 36 and 44 extended families (mean = 39), depending on the marker. There were an average of 7.9 affected subjects per extended family (see Bureau et al. (2013) and references therein for details of study sample, phenotype definition and genotyping procedures). Table 4 gives the sample mean and SD of the variance estimates over markers grouped by intervals of minor allele frequency. As in our simulations, the means of the model-based and robust variance estimates are close to each other. In the allele frequency intervals 0.2–0.35 and 0.35–0.5 with the most markers, the robust estimator is slightly more variable than the model-based estimator, but the difference is small given there are only on average 2 nuclear families per extended family.

4 Discussion

Subjects in different nuclear families connected to the same extended family depend on each other in genotype and phenotype. This has consequences on statistical inference when a risk locus influences the genotype transmitted at a nearby tested genetic marker (dependence due

to linkage) or the family-specific environment influences the effect of a risk variant. We investigated the common practice of estimating the effect size of a genetic variant using a within-nuclear-family analysis on a sample of nuclear families obtained by splitting extended families. With large sample theory and a single affected child per nuclear family, we have shown that the naive model-based estimate of the variance of the maximum conditional likelihood estimate of a logistic regression coefficient underestimates the true variance of the estimate when there is dependence among nuclear families within the same extended family. Simulations and numerical calculations however revealed that this underestimation is negligible when dependence of marker genotypes among cousins is due to a tightly linked risk locus with an allelic odds ratio of 1.5, to a substantial heterogeneity of the effect of a risk allele between extended families, with odds ratios ranging from 1 to 2.25, or to both sources of dependence simultaneously. When testing a variant unassociated to the disease, the variance underestimation was insufficient to provoke an inflation of the Type I error of the Wald test. When investigating the effect of genetic linkage on conditional logistic regression maximum likelihood inference in sibships, Siegmund et al. (2000) likewise found no substantial Type I error inflation of the Wald test even with an odds ratio as extreme as 20.

An empirical estimator of variance can be constructed to be robust to dependence among nuclear families from the same extended family. This protection from bias comes with a price in precision of the variance estimator, with the SD 2 to 4 times larger under our simulation scenario. This can increase Type I error, although the impact was very small in our simulations. While recognizing that the Wald test using the model-based estimator of variance may perform adequately for practical purposes Siegmund et al. (2000) argued for the use of the empirical estimator of variance for conditional logistic regression coefficient Wald tests in sibships to protect the Type I error. These authors found little impact on power compared to the test with the model-based estimator when testing the actual risk variant, an observation we replicated in our simulations with extended families. To the contrary, for an association analysis conditional on parental genotypes in nuclear families from extended families, we recommend using the model-based estimator given its greater precision compared to the empirical variance estimator. This recommendation is restricted to within-family association analysis of extended families, as other settings, such as spatial data with larger number of strata within clusters may lead to important underestimation of the variance of the conditional logistic regression coefficient estimates and a requirement to use the empirical variance estimate. This was the case in an analysis of the spatial distribution of plains bison where each animal is a cluster with multiple strata of observations (Craiu et al., 2008).

This article focuses on estimation of odds ratios (approximating relative risks) as measures of the effect size of genetic variants, which is currently only available with nuclear families. If one focuses instead solely on testing association, score tests such as the GDT developed for extended families were already shown to achieve greater power than score tests applied to nuclear families from extended families (Chen et al., 2009), and we observed a similar moderate power advantage of the GDT over Wald tests for nuclear families.

In our illustration, we discarded the nuclear families with missing parental genotypes, as the conditional likelihood of equation 2 requires complete parental genotype data. Various strategies can be applied to deal with missing parental genotypes in extended families where the genotype of multiple distant relatives can provide information to infer the distribution of missing genotypes. It is possible to infer the genotype distribution of parents given all observed genotypes in extended families using pedigree likelihoods implemented in software packages such as Merlin (csg.sph.umich.edu/abecasis/merlin/), but these require the restrictive assumptions of random mating and Hardy-Weinberg equilibrium on genotype distributions, such that analyses where possible parental genotypes are weighted according to these inferred genotype distributions may be sensitive to population stratification. A number of approaches partially robust to population stratification have been proposed to handle missing parental genotypes in nuclear families (reviewed by Dudbridge (2008)). These can be applied to samples of nuclear families obtained from splitting extended families, but then the information from the genotypes of distant relatives is lost. Investigating the relative merits of these two approaches in terms of bias and power is beyond the scope of this paper.

In conclusion, the theory suggests that robust variance estimators should be used for Wald-type inferences under the common practice of splitting extended families into nuclear families and deriving the score estimating equations by treating nuclear families as if they were independent, because the model-based variance estimator will tend to underestimate the true variance of the coefficient estimates. However in finite samples of practical size and under realistic assumptions for the mean allelic odds ratio and its variation from family to family, standard model-based methods lead to valid and more efficient statistical inferences, even in presence of genetic linkage to a risk locus or effect heterogeneity of a magnitude that can be realistically expected in association studies of common genetic variants.

Acknowledgments

Financial support from the Canadian Institutes of Health research (CIHR, grant MOP-106448) (AB) and the Natural Sciences and Engineering Research Council of Canada (TD) is gratefully acknowledged. A. Bureau was supported by a research fellowship from the Fonds de recherche du Québec - Santé. We thank J. Croteau (Centre de Recherche de l'Institut Universitaire en Santé Mentale de Québec) for his programming assistance and M. Maziade for access to the data from the Eastern Quebec Schizophrenia and Bipolar Disorder Study, funded by CIHR (grants MOP-74430, MOP-119408 and MOP-114988) and by a Canada Research Chair (950-200810) in the genetics of neuropsychiatric disorders of which M. Maziade is the Chair.

References

- Allen-Brady K, Wong J, Camp NJ. PedGenie: an analysis approach for genetic association testing in extended pedigrees and genealogies of arbitrary size. *BMC Bioinformatics*. 2006; 7:209. [PubMed: 16620382]
- Arbogast PG, Lin DY. Goodness-of-fit methods for matched case-control studies. *Canad J Statist*. 2004; 32:373–386.
- Bureau A, Chagnon YC, Croteau J, Fournier A, Roy MA, Paccalet T, Merette C, Maziade M. Follow-up of a major psychosis linkage site in 13q13-q14 reveals significant association in both case-control and family samples. *Biol Psychiatry*. 2013; 74:444–50. [PubMed: 23602252]
- Bureau A, Croteau J, Chagnon YC, Roy MA, Maziade M. Extension of the generalized disequilibrium test to polytomous phenotypes and two-locus models. *Frontiers in Genetics*. 2014; 5:258. [PubMed: 25152751]

- Chen WM, Manichaikul A, Rich SS. A generalized family-based association test for dichotomous traits. *American Journal of Human Genetics*. 2009; 85:364–76. [PubMed: 19732865]
- Craiu RV, Duchesne T, Fortin D. Inference methods for the conditional logistic regression model with longitudinal data. *Biometrical Journal*. 2008; 50:97–109. [PubMed: 17849385]
- Dudbridge F. Likelihood-based association analysis for nuclear families and unrelated subjects with missing genotype data. *Human Heredity*. 2008; 66:87–98. [PubMed: 18382088]
- Dudbridge F, Holmans PA, Wilson SG. A flexible model for association analysis in sibships with missing genotype data. *Annals of Human Genetics*. 2011; 75:428–38. [PubMed: 21241274]
- Lake SL, Blacker D, Laird NM. Family-based tests of association in the presence of linkage. *American Journal of Human Genetics*. 2000; 67:1515–25. [PubMed: 11058432]
- Lange C, DeMeo D, Silverman EK, Weiss ST, Laird NM. PBAT: Tools for family-based association studies. *American Journal of Human Genetics*. 2004; 74:367–369. [PubMed: 14740322]
- Lynch, M., Walsh, B. *Genetics and Analysis of Quantitative Traits*. Sunderland, MA: Sinauer Associates; 1998.
- Martin ER, Bass MP, Hauser ER, Kaplan NL. Accounting for linkage in family-based tests of association with missing parental genotypes. *American Journal of Human Genetics*. 2003; 73:1016–26. [PubMed: 14551902]
- Martin ER, Bass MP, Kaplan NL. Correcting for a potential bias in the pedigree disequilibrium test. *American Journal of Human Genetics*. 2001; 68:1065–7. [PubMed: 11254459]
- Martin ER, Kaplan NL, Weir BS. Tests for linkage and association in nuclear families. *American Journal of Human Genetics*. 1997; 61:439–48. [PubMed: 9311750]
- McNeil, AJ., Frey, R., Embrechts, P. *Quantitative Risk Management. Concepts, Techniques and Tools*, Princeton Series in Finance. Princeton, NJ: Princeton University Press; 2005.
- Neuhaus JM, Jewell NP. The effect of retrospective sampling on binary regression models for clustered data. *Biometrics*. 1990; 46:977–90. [PubMed: 2085642]
- Self SG, Longton G, Kopecky KJ, Liang KY. On estimating hla/disease association with application to a study of aplastic anemia. *Biometrics*. 1991; 47:53–61. [PubMed: 2049513]
- Siegmund KD, Langholz B, Kraft P, Thomas DC. Testing linkage disequilibrium in sibships. *American Journal of Human Genetics*. 2000; 67:244–8. [PubMed: 10831398]
- Witte JS, Gauderman WJ, Thomas DC. Asymptotic bias and efficiency in case-control studies of candidate genes and gene-environment interactions: basic family designs. *American Journal of Epidemiology*. 1999; 149:693–705. [PubMed: 10206618]
- Ziegler, A., RKönig, I. *A statistical approach to genetic epidemiology*. 2. Weinheim: Wiley-VCH; 2010.

Appendix: Additional derivations

A Variance of $\hat{\beta}$

In this section, we adopt a more general notation by introducing the index i for the observations within a nuclear family, which consist of the actual affected child and his pseudo-sibs created from the untransmitted parental genotypes in the analysis conditional on parental genotypes. This notation allows us to encompass the more general setting of $n_{CS} - 1$ control observations matched to a case observation, where conditional logistic regression is applicable.

We consider the special case with a single covariate. Let X_{CS}^{case} denote the value of the covariate for the case in stratum (nuclear family) s in cluster (extended family) c . The estimator $\hat{\beta}^*$ is the value of β that solves

$$U(\beta) = \frac{1}{K} \sum_c \sum_s \left\{ X_{cs}^{\text{case}} - \sum_{i=1}^{n_{cs}} X_{csi} \frac{\exp(\beta X_{csi})}{\sum_{i'=1}^{n_{cs}} \exp(\beta X_{csi'})} \right\} = 0.$$

We first compute

$$\Delta = \text{var}\{\sqrt{K}U(\beta)\} - \text{var}_{\text{MB}}\{\sqrt{K}U(\beta)\},$$

where $\text{var}()$ denotes the true variance and $\text{var}_{\text{MB}}()$ denotes the variance under the assumption that individuals from different nuclear families are independent. Let $E_{X_c}^{\text{1case}}$ and $\text{var}_{X_c}^{\text{1case}}$ respectively denote the mean and variance with respect to the distribution of the covariates when sampling nuclear families with one case in the c -th extended family. Then by independence of the extended families we get

$$\text{var}\{\sqrt{K}U(\beta^*)\} = K^{-1} \sum_c \text{var} \left[\sum_s \left\{ X_{cs}^{\text{case}} - \sum_{i=1}^{n_{cs}} X_{csi} \frac{\exp(\beta X_{csi})}{\sum_{i'=1}^{n_{cs}} \exp(\beta X_{csi'})} \right\} \right].$$

But

$$\begin{aligned} & \text{var} \left[\sum_s \left\{ X_{cs}^{\text{case}} - \sum_{i=1}^{n_{cs}} X_{csi} \frac{\exp(\beta X_{csi})}{\sum_{i'=1}^{n_{cs}} \exp(\beta X_{csi'})} \right\} \right] \\ &= E_{X_c}^{\text{1case}} \left(\text{var} \left[\sum_s \left\{ X_{cs}^{\text{case}} - \sum_{i=1}^{n_{cs}} X_{csi} \frac{\exp(\beta X_{csi})}{\sum_{i'=1}^{n_{cs}} \exp(\beta X_{csi'})} \right\} \middle| X_c, 1\text{case} \right] \right) \\ &+ \text{var}_{X_c}^{\text{1case}} \left(E \left[\sum_s \left\{ X_{cs}^{\text{case}} - \sum_{i=1}^{n_{cs}} X_{csi} \frac{\exp(\beta X_{csi})}{\sum_{i'=1}^{n_{cs}} \exp(\beta X_{csi'})} \right\} \middle| X_c, 1\text{case} \right] \right). \end{aligned} \quad (13)$$

The conditional expectation inside $\text{var}_{X_c}^{\text{1case}}$ on the RHS of (13) is zero. Let us calculate the conditional variance inside of $E_{X_c}^{\text{1case}}$ in (13), which we denote V for convenience. We now use the well known fact (Arbogast and Lin, 2004) that treating the subject with $Y=1$ as random given that there is exactly one such subject ($\sum_{i=1}^{n_{cs}} Y_{csi}=1$), keeping the values of X fixed, is equivalent to treating the X 's as random, given the set of observed X values and the Y s. This equivalence extends to the analysis conditional on parental genotypes, where assigning $Y=1$ randomly to one of the possible child genotypes formed from the parental genotypes is equivalent to the random transmission of a genotype to the affected child from the parental genotypes, under the actual disease risk model. Hence calculating V is equivalent to computing the following variance with respect to the distribution of $Y_{cs1}, \dots, Y_{csn_{cs}}$ given $\sum_{i=1}^{n_{cs}} Y_{csi}=1$ and X_{cs} :

$$\begin{aligned}
V &= \text{var} \left(\sum_{s=1}^{n_c} \sum_{i=1}^{n_{cs}} Y_{csi} X_{csi} - \sum_{i=1}^{n_{cs}} X_{csi} \frac{\exp(\beta X_{csi})}{\sum_{i'=1}^{n_{cs}} \exp(\beta X_{csi'})} \middle| \sum_{i=1}^{n_{cs}} Y_{csi}=1, X_c \right) \\
&= \text{var} \left(\sum_{s=1}^{n_c} \sum_{i=1}^{n_{cs}} Y_{csi} X_{csi} \middle| \sum_{i=1}^{n_{cs}} Y_{csi}=1, X_c \right) \\
&= \sum_{s=1}^{n_c} \sum_{i=1}^{n_{cs}} X_{csi}^2 \text{var} \left(Y_{csi} \middle| \sum_{i=1}^{n_{cs}} Y_{csi}=1, X_c \right) + 2 \sum_s \sum_{s'} \sum_i \sum_{i'} X_{csi} X_{cs'i'} \text{cov} \left(Y_{csi}, Y_{cs'i'} \middle| \sum_{i=1}^{n_{cs}} Y_{csi}=1, X_c \right).
\end{aligned}$$

The only difference between the naive model-based variance and the true variance is that the former falsely assumes that $\text{cov}(Y_{csi}, Y_{cs'i'} | \sum_i Y_{csi}=1, X_c) = 0$ when $s \neq s'$. Hence we have that

$$\Delta = E_{X_c}^{\text{1case}} \left\{ \frac{2}{K} \sum_c \sum_s \sum_{s'} \sum_x \sum_{x'} n_{xx'}^{css'} x x' \sigma_{xx'}^{css'} \right\}, \quad (14)$$

where $n_{xx'}^{css'}$ is the number of terms in $\sum_x \sum_{x'} X_{csi}=x$ and $X_{cs'i'}=x'$ and

$$\sigma_{xx'}^{css'} = \text{cov} \left(Y_{csi}, Y_{cs'i'} \middle| \sum_i Y_{csi}=1, X_{csi}=x, X_{cs'i'}=x', X_c \right). \quad (15)$$

Now

$$\begin{aligned}
\text{var}(\sqrt{K} \hat{\beta}^*) &\simeq \frac{\text{var}\{\sqrt{K} U(\beta^*)\}}{[E\{\frac{\partial}{\partial \beta^*} \sqrt{K} U(\beta^*)\}]^2} = \frac{\text{var}_{\text{MB}}\{\sqrt{K} U(\beta^*)\} + \Delta}{[E\{\frac{\partial}{\partial \beta^*} \sqrt{K} U(\beta^*)\}]^2} \\
&= \text{var}_{\text{MB}}(\sqrt{K} \hat{\beta}^*) + \Delta \{\text{var}_{\text{MB}}(\sqrt{K} \hat{\beta}^*)\}^2 \\
&= \text{var}_{\text{MB}}(\sqrt{K} \hat{\beta}^*) \{1 + \Delta \text{var}_{\text{MB}}(\sqrt{K} \hat{\beta}^*)\} \\
\Rightarrow \text{var}(\hat{\beta}^*) &\simeq \text{var}_{\text{MB}}(\hat{\beta}^*) \{1 + K \Delta \text{var}_{\text{MB}}(\hat{\beta}^*)\}, \quad (16)
\end{aligned}$$

which is the variance inflation formula given by equation (2.8) in the main article.

B Sign of

We now determine the sign of Δ in certain special cases where the covariate X_{csi} is discrete and where the stratum specific random effects can be written as $\gamma_{cs}^{(j)} = \tilde{\gamma}_c^{(j)} + \tilde{\gamma}_{cs'}^{(j)}$, $j = 0, 1$, with $\tilde{\gamma}_c^{(j)}, \tilde{\gamma}_{cs'}^{(j)}, \tilde{\gamma}_{cs}^{(j)}$ independent; in other words under an exchangeable correlation structure such that $\text{corr}(Y_{cs}^{(j)}, Y_{cs'}^{(j)}) = r_c^{(j)}$. (These conditions will allow us to invoke a comonotonicity argument in sub-sections B.1 and B.2.) For instance, in the genetic context, the phase between a risk variant and a marker allele is common to whole extended families (see Section C below), creating a constant correlation among the alleles transmitted to the

children of all nuclear families within an extended family at the risk variant site. In the remainder of this section γ_c denotes the vector of all random effects in cluster c .

We must first compute $\sigma_{xx'}^{css'}$ given by (15). We have that

$$\begin{aligned}\sigma_{xx'}^{css'} &= \int \Pr(Y_{csi}=1, Y_{cs'i'}=1 | \gamma_c, X_c, \sum Y=1) dF_{\gamma|c}(\gamma) \\ &\quad - \int \Pr(Y_{csi}=1 | \gamma_c, X_c, \sum Y=1) dF_{\gamma|c}(\gamma) \\ &\quad \times \int \Pr(Y_{cs'i'}=1 | \gamma_c, X_c, \sum Y=1) dF_{\gamma|c}(\gamma),\end{aligned}\quad (17)$$

where $F_{\gamma|c}$ denotes the joint cdf of the random effects in cluster c given the observed covariates and stratum sums in cluster c . Let n_x^{cs} denote the number of observations in nuclear family s of extended family c that have $X_{csi}=x$. Conditionally on the random effects, observations in the nuclear families are independent and follow the logistic regression model, yielding

$$\Pr(Y_{csi}=1 | \gamma_c, X_c, \sum Y=1) = \frac{e^{(\beta + \tilde{\gamma}_c^{(1)} + \tilde{\gamma}_{cs}^{(1)})x}}{n_x^{cs} e^{(\beta + \tilde{\gamma}_c^{(1)} + \tilde{\gamma}_{cs}^{(1)})x} + n_x^{cs'} e^{(\beta + \tilde{\gamma}_c^{(1)} + \tilde{\gamma}_{cs'}^{(1)})x'} + (n^{cs} - n_x^{cs} - n_x^{cs'})}. \quad (18)$$

We can exchange s and s' in (18) to obtain $\Pr(Y_{cs'i'}=1 | \gamma_c, X_c, \sum Y=1)$.

Finally because of the conditional independence of the Y_{csi} 's given the random effects, we have that

$$\begin{aligned}\Pr(Y_{csi}=1, Y_{cs'i'}=1 | \gamma_c, X_c, \sum Y=1) &= \Pr(Y_{csi}=1 | \gamma_c, X_c, \sum Y=1) \\ &\quad \times \Pr(Y_{cs'i'}=1 | \gamma_c, X_c, \sum Y=1).\end{aligned}\quad (19)$$

Define

$$\begin{aligned}U_s(\tilde{\gamma}_c^{(1)}; \tilde{\gamma}_{cs}^{(1)}) &= \frac{e^{(\beta + \tilde{\gamma}_c^{(1)} + \tilde{\gamma}_{cs}^{(1)})x}}{n_x^{cs} e^{(\beta + \tilde{\gamma}_c^{(1)} + \tilde{\gamma}_{cs}^{(1)})x} + n_x^{cs'} e^{(\beta + \tilde{\gamma}_c^{(1)} + \tilde{\gamma}_{cs'}^{(1)})x'} + (n^{cs} - n_x^{cs} - n_x^{cs'})} \\ \text{and } U_{s'}(\tilde{\gamma}_c^{(1)}; \tilde{\gamma}_{cs'}^{(1)}) &= \frac{e^{(\beta + \tilde{\gamma}_c^{(1)} + \tilde{\gamma}_{cs'}^{(1)})x'}}{n_x^{cs'} e^{(\beta + \tilde{\gamma}_c^{(1)} + \tilde{\gamma}_{cs'}^{(1)})x'} + n_x^{cs} e^{(\beta + \tilde{\gamma}_c^{(1)} + \tilde{\gamma}_{cs}^{(1)})x} + (n^{cs'} - n_x^{cs'} - n_x^{cs})}.\end{aligned}$$

Then from (17), (18) and (19), we have that $\sigma_{xx'}^{css'} = \text{cov}\{U_s(\tilde{\gamma}_c^{(1)}; \tilde{\gamma}_{cs}^{(1)}), U_{s'}(\tilde{\gamma}_c^{(1)}; \tilde{\gamma}_{cs'}^{(1)})\}$.

B.1 Case of a binary X_{csi}

Without loss of generality, assume that X_{csi} may only take on values 0 or 1 (any binary X_{csi} can be shifted and rescaled without changing the logistic regression model by suitably shifting and rescaling the α and β coefficients). In this case, the product xx' is either equal to 0 or 1. Hence the only terms contributing to are those for which $x = x' = 1$. Note that in this case both $U_s(\tilde{\gamma}_c^{(1)}; \tilde{\gamma}_{cs}^{(1)})$ and $V_{s'}(\tilde{\gamma}_c^{(1)}; \tilde{\gamma}_{cs'}^{(1)})$ are increasing functions of $\tilde{\gamma}_c^{(1)}$ when $\tilde{\gamma}_{cs}^{(1)}$ and $\tilde{\gamma}_{cs'}^{(1)}$ are fixed and are thus comonotonic. By the Höfdding formula for the covariance (see McNeil, Frey, and Embrechts (2005), Lemma 5.24), we get that

$$\text{cov}\{U_s(\tilde{\gamma}_c^{(1)}; \tilde{\gamma}_{cs}^{(1)}), V_{s'}(\tilde{\gamma}_c^{(1)}; \tilde{\gamma}_{cs'}^{(1)}) | \tilde{\gamma}_{cs}^{(1)}, \tilde{\gamma}_{cs'}^{(1)}\} \geq 0, \forall \tilde{\gamma}_{cs}^{(1)}, \tilde{\gamma}_{cs'}^{(1)} \quad (20)$$

when $x = x' = 1$. But

$$\begin{aligned} \text{cov}\{U_s(\tilde{\gamma}_c^{(1)}; \tilde{\gamma}_{cs}^{(1)}), V_{s'}(\tilde{\gamma}_c^{(1)}; \tilde{\gamma}_{cs'}^{(1)})\} = & E \left[\text{cov}\{U_s(\tilde{\gamma}_c^{(1)}; \tilde{\gamma}_{cs}^{(1)}), V_{s'}(\tilde{\gamma}_c^{(1)}; \tilde{\gamma}_{cs'}^{(1)}) | \tilde{\gamma}_{cs}^{(1)}, \tilde{\gamma}_{cs'}^{(1)}\} \right] \\ & + \text{cov} \left[E\{U_s(\tilde{\gamma}_c^{(1)}; \tilde{\gamma}_{cs}^{(1)}) | \tilde{\gamma}_{cs}^{(1)}\}, E\{V_{s'}(\tilde{\gamma}_c^{(1)}; \tilde{\gamma}_{cs'}^{(1)}) | \tilde{\gamma}_{cs'}^{(1)}\} \right]. \end{aligned} \quad (21)$$

By (20), we have that the first term on the RHS of (21) is non-negative. As for the second term, it is of the form $\text{cov}[f_1(\tilde{\gamma}_{cs}^{(1)}), f_2(\tilde{\gamma}_{cs'}^{(1)})]$, which is zero by independence of $\tilde{\gamma}_{cs}^{(1)}$ and $\tilde{\gamma}_{cs'}^{(1)}$. Hence we have that 0 when X_{csi} is binary.

B.2 Case of X_{csi} taking on three equidistant values $k_1, k_1 + k_2, k_1 + 2k_2$

Again without loss of generality, assume that X_{csi} takes on a value among $-1, 0$ or 1 after dividing by k_2 and subtracting $\frac{k_1}{k_2} + 1$. The terms that contribute to are those for which neither x nor x' take on the value 0. When $x = 1$ and $x' = -1$ and $\tilde{\gamma}_{cs}^{(1)}$ and $\tilde{\gamma}_{cs'}^{(1)}$ are fixed, then $U_s(\tilde{\gamma}_c^{(1)}; \tilde{\gamma}_{cs}^{(1)})$ and $V_{s'}(\tilde{\gamma}_c^{(1)}; \tilde{\gamma}_{cs'}^{(1)})$ are respectively increasing and decreasing functions of $\tilde{\gamma}_c^{(1)}$ and are thus anti-monotonic. In this case the Höfdding formula for the covariance leads to $\text{cov}\{U_s(\tilde{\gamma}_c^{(1)}; \tilde{\gamma}_{cs}^{(1)}), V_{s'}(\tilde{\gamma}_c^{(1)}; \tilde{\gamma}_{cs'}^{(1)})\} \leq 0$; the same result holds when $x = -1$ and $x' = 1$. On the other hand when x and x' are either both 1 or both -1 , then $U_s(\tilde{\gamma}_c^{(1)}; \tilde{\gamma}_{cs}^{(1)})$ and $V_{s'}(\tilde{\gamma}_c^{(1)}; \tilde{\gamma}_{cs'}^{(1)})$ are comonotonic for fixed $\tilde{\gamma}_{cs}^{(1)}$ and $\tilde{\gamma}_{cs'}^{(1)}$, and so $\text{cov}\{U_s(\tilde{\gamma}_c^{(1)}; \tilde{\gamma}_{cs}^{(1)}), V_{s'}(\tilde{\gamma}_c^{(1)}; \tilde{\gamma}_{cs'}^{(1)})\} \geq 0$. Therefore, $\sigma_{xx'}^{css'}$ is non-negative when $xx' = 1$ and is non-positive when $xx' = -1$, which implies that 0.

C Between-nuclear-family covariance due to genetic linkage

We present in this section the approximate computation of (15) for a pair of affected first cousins when dependence between their marker M genotypes is due to genetic linkage to a nearby risk locus. In the notation from the main text, we have

$$\sigma_{xx'}^{cs'} = \Pr \left(Y_{cs}=1, Y_{cs'}=1 \mid \sum_{h \in S_{cs}} Y_h=1, X_{cs}=x, \sum_{h' \in S_{cs'}} Y_{h'}=1, X_{cs'}=x', X_c \right) \\ - \Pr \left(Y_{cs}=1 \mid \sum_{h \in S_{cs}} Y_h=1, X_{cs}=x, X_c \right) \Pr \left(Y_{cs'}=1 \mid \sum_{h' \in S_{cs'}} Y_{h'}=1, X_{cs'}=x', X_c \right)$$

where Y_h denotes an indicator for whether genotype h was transmitted or not.

Now, introduce in the joint probability of Y_{cs} and $Y_{cs'}$ the risk variant genotype G coded by the number of minor alleles V and the indicator ϕ_c of the phase between the minor alleles of G and M in the extended families (1 = in cis, 0 = in trans):

$$\Pr \left(Y_{cs}=1, Y_{cs'}=1 \mid \sum_{h \in S_{cs}} Y_h=1, X_{cs}=x, \sum_{h' \in S_{cs'}} Y_{h'}=1, X_{cs'}=x', X_c \right) \\ = \sum_{\phi_c} \sum_{V_c} \Pr \left(Y_{cs}=1, Y_{cs'}=1, V_c, \phi_c \mid \sum_{h \in S_{cs}} Y_h=1, X_{cs}=x, \sum_{h' \in S_{cs'}} Y_{h'}=1, X_{cs'}=x', X_c \right) \\ = \sum_{\phi_c} \left(\sum_{V_{cs}} \sum_{V_{cs'}^{(p)}} \sum_{V_{cs'}^{(f)}} \left[\Pr \left(Y_{cs}=1 \mid \sum_{h \in S_{cs}} Y_h=1, V_{cs}, V_{cs'}^{(p)}, V_{cs'}^{(f)} \right) \right. \right. \\ \times \Pr \left(V_{cs} \mid X_{cs}=x, V_{cs}^{(p)}, V_{cs}^{(f)}, X_{cs}^{(p)}, X_{cs}^{(f)}, \phi_c \right) \Pr \left(V_{cs'}^{(f)} \mid X_{cs'}^{(f)} \right) \\ \times \left\{ \sum_{V_{cs'}} \sum_{V_{cs'}^{(p)}} \sum_{V_{cs'}^{(f)}} \Pr \left(Y_{cs'}=1 \mid \sum_{h' \in S_{cs'}} Y_{h'}=1, V_{cs'}, V_{cs'}^{(p)}, V_{cs'}^{(f)} \right) \right. \\ \times \Pr \left(V_{cs'} \mid X_{cs'}=x', V_{cs'}^{(p)}, V_{cs'}^{(f)}, X_{cs'}^{(p)}, X_{cs'}^{(f)}, \phi_c \right) \Pr \left(V_{cs'}^{(f)} \mid X_{cs'}^{(f)} \right) \\ \left. \left. \times \Pr \left(V_{cs}^{(p)}, V_{cs'}^{(p)}, \phi_c \mid X_{cs}^{(p)}, X_{cs'}^{(p)} \right) \right\} \right] \right),$$

where the superscript (p) refers to the parent of a nuclear family who is a sibling of the parent of the other nuclear family, i.e. who connects the nuclear family s to the extended family c and the superscript (f) refers to the other parent (married-in parent, who can be considered founder of the pedigree). As an example, Table 5 gives genotype probabilities $\Pr(V \mid X, V^{(p)}, V^{(f)}, X^{(p)}, X^{(f)}, \phi)$ when $X^{(p)} = V^{(p)} = 1$ and $X^{(f)} = V^{(f)} = 0$ as a function of the

phase indicator ϕ . The terms $\Pr \left(V_{cs}^{(p)}, V_{cs'}^{(p)}, \phi_c \mid X_{cs}^{(p)}, X_{cs'}^{(p)} \right)$ are functions of the risk variant frequency p_v , the minor allele frequency of the marker p_x and the coefficient of linkage disequilibrium δ between the two in the population. Linkage disequilibrium is usually

expressed by the squared correlation r^2 , and we use the equation $\delta = \sqrt{r^2 p_x (1-p_x) p_v (1-p_v)}$ to obtain δ .

The terms $\Pr \left(Y_{cs}=1 \mid \sum_{h \in S_{cs}} Y_h=1, V_{cs}, V_{cs'}^{(p)}, V_{cs'}^{(f)} \right)$ are standard conditional logistic probabilities, with $V_{cs}^{(p)}, V_{cs'}^{(f)}$ defining the set of possible genotypes S_{cs} :

$$\Pr \left(Y_{cs}=1 \mid \sum_{h \in S_{cs}} Y_h=1, V_{cs}, V_{cs}^{(p)}, V_{cs}^{(f)} \right) = \frac{\exp(\beta V_{cs})}{\sum_{h^* \in S_{cs}} \exp(\beta V_{cs}^{(h^*)})}.$$

As an approximation, we assumed only the connecting parent could carry a risk allele, i.e. $V_{cs}^{(f)}=0$, and considered only $X=0, 1$, neglecting the probability that $X=2$. We only computed this approximation with allele frequencies up to 0.1 to insure accuracy. Restricting ourselves to low allele frequencies made sense since formula (14) is only valid when $\Pr(Y_{cs}=1 \mid X_{cs}=x)$ follows a logistic model, and in the current setting where X is the allele count of a marker and not a causal variant this is true only when X is dichotomous (see Section 2.2.1 of the main text).

Once the joint probabilities have been obtained, the probabilities in each nuclear family $\Pr(Y_{cs}=1 \mid \sum_{h \in S_{cs}} Y_h=1, X_{cs}=x, X_c)$ are obtained by summing over the outcomes in the other nuclear family.

We also need the probabilities of observing the values of the pair (X_{cs}, X_{cs}') . We need only

$\Pr \left(X_{cs}=X_{cs}'=1 \mid \sum_{h \in S_{cs}} Y_h=1, \sum_{h' \in S_{cs}'} Y_{h'}=1 \right)$, since we assume $X=0, 1$. This probability equals the expectation of the term xx' in (14). It is computed by introducing the number of minor alleles V of the risk variant genotype of the children and their parents, along with the phase indicator ϕ_c , as for the computation of $\sigma_{xx'}^{css'}$.

Values of $\sigma_{xx'}^{css'}$ and the expectation of xx' computed for a set of values of p_v, p_x and δ can then be plugged into formulae (14) and (16) to obtain an approximation of the true variance.

Table 1

Variance estimation of the logistic regression coefficient estimate

	Allele frequency		
	0.1	0.3	0.5
<i>Fixed effect</i>			
Risk allele			
Mean of model-based variance estimate (SD)	0.0281 (0.0027)	0.0131 (0.0008)	0.0120 (0.0006)
Mean of robust variance estimate (SD)	0.0275 (0.0058)	0.0128 (0.0024)	0.0118 (0.0022)
Sample variance of $\hat{\beta}$ [95% CI] ^a	0.0282 [0.0271, 0.0293]	0.0131 [0.0126, 0.0136]	0.0118 [0.0113, 0.0123]
Allele strongly associated to risk allele ($r^2 = 0.8$)			
Mean of model-based variance estimate (SD)	0.0283 (0.0027)	0.0131 (0.0007)	0.0119 (0.0006)
Mean of robust variance estimate (SD)	0.0278 (0.0058)	0.0129 (0.0024)	0.0117 (0.0022)
Predicted variance of $\hat{\beta}$	0.0295	NA	NA
Sample variance of $\hat{\beta}$ [95% CI] ^a	0.0288 [0.0277, 0.0299]	0.0132 [0.0127, 0.0137]	0.0119 [0.0114, 0.0124]
Allele unassociated to risk allele ($r^2 = 0$)			
Mean of model-based variance estimate (SD)	0.0323 (0.0032)	0.0137 (0.0008)	0.0115 (0.0005)
Mean of robust variance estimate (SD)	0.0318 (0.0067)	0.0135 (0.0025)	0.0113 (0.0021)
Predicted variance of $\hat{\beta}$	0.0330	NA	NA
Sample variance of $\hat{\beta}$ [95% CI] ^a	0.0323 [0.0310, 0.0336]	0.0140 [0.0135, 0.0145]	0.0115 [0.0111, 0.0119]
<i>Extended family-specific effect due to shared environment</i>			
Risk allele			
Mean of model-based variance estimate (SD)	0.0251 (0.0025)	0.0116 (0.0008)	0.0105 (0.0006)
Mean of robust variance estimate (SD)	0.0254 (0.0053)	0.0120 (0.0023)	0.0107 (0.0021)
Sample variance of $\hat{\beta}$ [95% CI] ^a	0.0261 [0.0251, 0.0271]	0.0123 [0.0118, 0.0128]	0.0112 [0.0108, 0.0116]
Allele strongly associated to risk allele ($r^2 = 0.8$)			
Mean of model-based variance estimate (SD)	0.0254 (0.0025)	0.0115 (0.0007)	0.0104 (0.0006)
Mean of robust variance estimate (SD)	0.0257 (0.0053)	0.0120 (0.0023)	0.0105 (0.0020)
Sample variance of $\hat{\beta}$ [95% CI] ^a	0.0260 [0.0250, 0.0270]	0.0122 [0.0117, 0.0127]	0.0110 [0.0106, 0.0114]

	Allele frequency			
	0.1	0.3	0.5	
Allele unassociated to risk allele ($r^2 = 0$)	Mean of model-based variance estimate (SD)	0.0285 (0.0028)	0.0120 (0.0007)	0.0099 (0.0006)
	Mean of robust variance estimate (SD)	0.0278 (0.0061)	0.0120 (0.0024)	0.0098 (0.0019)
	Sample variance of $\hat{\beta}$ [95% CI] ^a	0.0303	0.0117	0.0099
		[0.0291, 0.0315]	[0.0112, 0.0122]	[0.0095, 0.0103]

^aCI: Confidence Interval.

Table 2

Type I Error and Power

Allelic odds ratio	Test statistic	Variance estimator	Allele frequency		
			0.1	0.3	0.5
1	Wald	Model-based	0.049	0.050	0.049
		Robust	0.053	0.059	0.055
Fixed at 1.5	Wald	Model-based	0.692	0.950	0.964
		Robust	0.695	0.950	0.964
	GDT ^a	Kinship-based	0.790	0.983	0.990
Random with mean of 1.5	Wald	Model-based	0.802	0.990	0.997
		Robust	0.800	0.988	0.996
	GDT ^a	Kinship-based	0.900	0.999	1.000

^aGeneralized Disequilibrium Test

Table 395% Confidence Interval Coverage of Risk Allele and Strongly Associated Allele Coefficients^a

		Allele frequency		
	Variance estimator	0.1	0.3	0.5
<i>Fixed effect</i>				
Risk allele	Model-based	0.953	0.949	0.955
	Robust	0.946	0.943	0.948
Strongly associated allele ($r^2 = 0.8$)	Model-based	0.950	0.950	0.952
	Robust	0.945	0.941	0.943
<i>Extended family-specific effect due to shared environment</i>				
Risk allele	Model-based	0.948	0.944	0.938
	Robust	0.945	0.945	0.940
Strongly associated allele ($r^2 = 0.8$)	Model-based	0.950	0.944	0.940
	Robust	0.946	0.945	0.936

^aFor the associated allele and the risk allele with extended family-specific effect, the average true effect is unknown, and we estimated it using the mean of the coefficient estimates from the simulations.

Table 4

Variance estimation of the logistic regression coefficient estimate for markers in the Eastern Quebec Schizophrenia and Bipolar Disorder Study

	Minor allele frequency intervals		
	0.05 – 0.2	0.2 – 0.35	0.35 – 0.5
Number of markers	8	19	21
Mean of model-based variance estimate (SD)	0.20 (0.06)	0.13 (0.04)	0.099 (0.020)
Mean of robust variance estimate (SD)	0.18 (0.05)	0.13 (0.05)	0.094 (0.021)

Table 5

Example of a child's risk locus minor allele count V probabilities given his marker minor allele count X when $X^{(p)} = V^{(p)} = 1$ and $X^{(f)} = V^{(f)} = 0$

V		
X	0	1
$\phi = 1$		
0	1	0
1	0	1
$\phi = 0$		
0	0	1
1	1	0