When is an endophenotype useful to detect association to a disease? Exploring the relationships between disease status, endophenotype and genetic polymorphisms

Alexandre Bureau[1,2], Jordie Croteau[2]

Received _____; accepted _____

[1]Département de médecine sociale et préventive, Université Laval

[2]Institut universitaire en santé mentale de Québec du Centre intégré de santé et de services sociaux de la Capitale-Nationale

Appendix

1. Model selection in linear models of allele frequency

We retained from an exhaustive search the most parsimonious sub-model with a R^2 no more than 1 percent below the R^2 of the following two large models, one where the interaction between X_1 and Y_1 is the focus and the other where the interaction between X_1 and Y_2 is the focus.

$$E[X_2|X_1, Y_1, Y_2] = \beta_0 + \beta_1 X_1 + \beta_2 Y_1 + \beta_3 Y_2 + \beta_4 Y_1 Y_2 + \beta_5 X_1 Y_1 + \beta_6 X_1 Y_1 Y_2$$

$$E[X_2|X_1, Y_1, Y_2] = \beta_0 + \beta_1 X_1 + \beta_2 Y_1 + \beta_3 Y_2 + \beta_4 Y_1 Y_2 + \beta_5 X_1 Y_2 + \beta_6 X_1 Y_1 Y_2$$

In most instances, the SSE of the selected model did not exceed 1 percent of the total variance. For scenarios where no model satisfied this criterion, we retained the model with the lowest SSE and having at most three terms.

2. Simulating a bidimensional dichotomous phenotype in families

With a unidimensional dichotomous phenotype, dependance between relatives due to additive polygenic effects is usually modeled via a normally distributed latent variable u with correlation structure given by the kinship matrix of the relatives. The latent variable u is added to the fixed effects of genotypes at susceptibility loci in the mechanism, and the dichotomous status is obtained either by applying a threshold to the latent sum [1] or by transforming the sum into a phenotype probability via a link function in what is known as a generalized linear mixed model [2]. We adopt the latter approach.

With a bidimensional dichotomous phenotype (Y_1, Y_2), the cross-dependance between Y_1 and Y_2 within an individual as well as between relatives must also be included in

the simulation scheme. Given the nature of endophenotypes as intermediate phenotypes between genotype and disease status, we adopt a two-step approach: we first model the distribution of the vector Y_1, then the distribution of the vector $Y_2|Y_1$, in the spirit of transition models. (Note however that the fixed effects of the genotypes at the loci in the model can still be specified as a polytomous model, which is then reexpressed as a transition model, see for instance the supplementary material of [3]).

2.1. Simulation of Y_1 and genotypes

SIMLA was used to generate genotypes at two independent loci and at the same time the endophenotype Y_1. At each locus, one SNP was simulated which was perfectly linked to the disease locus and in perfect linkage disequilibrium with the disease-susceptibility alleles. The genotypes of pedigree founders were sampled under Hardy-Weinberg equilibrium using risk allele frequencies of 0.1 at locus 1 and 0.3 at locus 2. For the transition scenarios, simulation parameters could be obtained by computing marginal 2-locus penetrances for Y_1 and by setting both weights of the modes of inheritance equal to 0.5 (value corresponding to allelic mode). For polytomous scenarios, we had to approximate the simulation model by choosing the combination of weights of modes of inheritance and 2-locus penetrances that best fitted the Y_1 penetrance matrices of our scenarios.

2.2. Simulation of Y_2

In the model for $Y_{i2}|Y_1$ for subject i, Y_1 is treated as a vector of fixed effect, with the effect of the endophenotype of subject h, Y_{h1}, modulated by the kinship coefficient ϕ_{ih} between i and h, inspired by an additive polygenic effect. Adopting the logistic link

function, the model can be written:

$$\log\left(\frac{P[Y_{i2}=1|Y_1,X,U]}{P[Y_{i2}=0|Y_1,X,U]}\right) = \gamma'(X_i,Y_{i1}) + U_i + \alpha\sum_{h\neq i}^{n}(Y_{h1}-\nu)\phi_{ih} \quad (1)$$

$$U \sim N(0,\sigma^2\phi) \quad (2)$$

where ϕ is the kinship matrix between the family members and $\gamma'(X_i, Y_{i1})$ in an abreviated expression of the model for the disease phenotype given the genotype at major loci and endophenotype status of subject i (derived e.g. from a transition or polytomous model). In the case of a polytomous model, $\gamma'(X_i, Y_{i1}) = Y_{i1}(\beta_3 - \beta_1)X_i + (1 - Y_{i1})\beta_2 X_i$. The parameter σ^2 controls the degree of polygenic dependence between the disease status Y_2 of the family members and the parameter α the degree of genetic dependance of Y_2 on Y_1 not captured by the genotype at the loci in the model. The parameter ν, between 0 and 1, determines the relative importance of the risk increase $1-\nu$ due to observing an endophenotype impairment and the risk decrease $-\nu$ due to observing the normal level of the endophenotype in a relative.

It is important to note here that epidemiological studies are not usually designed to estimate disease prevalence in subjects with and without an endophenotype impairment and their relatives. The association is usually measured in reverse, by estimating prevalence of endophenotypes in disease patients, their non-affected relatives and unrelated controls. The invariance of the OR with respect to the sampling design can here be exploited, since the marginal ORs for Y_2 given Y_1 equals the OR of Y_1 given Y_2 estimated in epidemiological studies. Values of α and γ can then be selected to obtain in simulated data sets marginal ORs between Y_1 and Y_2 close to epidemiological estimates.

2.3. Family Ascertainment

At the step of simulation of Y_1, the SIMLA program selected families having at least the proband with the endophenotype impairment. Then among those families selected by SIMLA, we selected the ones having at least one cousin pair with the disease (Y_2). Given the high prevalence of the endophenotype impairment, especially among affected subjects, the selection step by SIMLA represents a negligible ascertainment on the endophenotype, the probability of having at least one subject with an endophenotype impairment being close to one in families having at least one affected cousin pair.

3. Proof of properties 10 and 13

Under the polytomous logistic model assume that the logistic function contrasting $Y_2 = 1$ and $Y_2 = 0$ when $Y_1 = 0$ is a general function $h(X_2, Z)$ of X_2 and a vector of other variables Z. Then

$$\begin{aligned}
& P[X_2 = u|Y_1 = 0, Y_2 = 1] \\
= & \frac{P[Y_1 = 0, Y_2 = 1|X_2 = u]P[X_2 = u]}{\sum_{u^*} P[Y_1 = 0, Y_2 = 1|X_2 = u^*]P[X_2 = u^*]} \\
= & \frac{\sum_z P[Y_1 = 0, Y_2 = 1|X_2 = u, Z = z]P[X_2 = u, Z = z]}{\sum_{u^*,z} P[Y_1 = 0, Y_2 = 1|X_2 = u^*, Z = z]P[X_2 = u^*, Z = z]} \\
= & \frac{\sum_z P[Y_1 = 0, Y_2 = 0|X_2 = u, Z = z] \exp(h(X_2 = u, Z = z))P[X_2 = u, Z = z]}{\sum_{u^*,z} P[Y_1 = 0, Y_2 = 0|X_2 = u^*, Z = z] \exp(h(X_2 = u^*, Z = z))P[X_2 = u^*, Z = z]}
\end{aligned}$$

We want the above expression to equal

$$P[X_2 = u|Y_1 = 0, Y_2 = 0] = \frac{P[Y_1 = 0, Y_2 = 0|X_2 = u]P[X_2 = u]}{\sum_{u^*} P[Y_1 = 0, Y_2 = 0|X_2 = u^*]P[X_2 = u^*]}$$

This is achieved only when $h()$ does not depend on X_2 and either $h()$ does not depend on Z neither, i.e. $h(X_2 = u, Z) = \beta_{20}$ a constant, or $h(X_2 = u, Z)$ depends on Z alone

$h(X_2 = u, Z) = g(Z)$ and $P[Z] = p$ a constant. Since $P[Z]$ is not in general equal for all values of Z, we require $h(X_2 = u, Z) = \beta_{20}$. This holds for all values of u.

If instead one considers the less stringent version of the property restricted to a fixed value of Z, then $h(X_2 = u, Z) = g(Z)$ is sufficient.

Under the transition model, we have

$$
\begin{aligned}
& P[X_2 = u | Y_1 = 0, Y_2 = 1] \\
= & \frac{P[Y_1 = 0, Y_2 = 1 | X_2 = u] P[X_2 = u]}{\sum_{u^*} P[Y_1 = 0, Y_2 = 1 | X_2 = u^*] P[X_2 = u^*]} \\
= & \frac{P[Y_2 = 1 | Y_1 = 0, X_2 = u] P[Y_1 = 0 | X_2 = u] P[X_2 = u]}{\sum_{u^*} P[Y_2 = 1 | Y_1 = 0, X_2 = u^*] P[Y_1 = 0 | X_2 = u^*] P[X_2 = u^*]} \\
= & \frac{\sum_z P[Y_2 = 1 | Y_1 = 0, X_2 = u, Z = z] P[Y_1 = 0 | X_2 = u, Z = z] P[X_2 = u, Z = z]}{\sum_{u^*, z} P[Y_2 = 1 | Y_1 = 0, X_2 = u^*, Z = z] P[Y_1 = 0 | X_2 = u, Z = z] P[X_2 = u^*, Z = z]}
\end{aligned}
$$

We want the above expression to equal

$$
\begin{aligned}
& P[X_2 = u | Y_1 = 0, Y_2 = 0] \\
= & \frac{P[Y_1 = 0, Y_2 = 0 | X_2 = u] P[X_2 = u]}{\sum_{u^*} P[Y_1 = 0, Y_2 = 0 | X_2 = u^*] P[X_2 = u^*]} \\
= & \frac{\sum_z P[Y_2 = 0 | Y_1 = 0, X_2 = u, Z = z] P[Y_1 = 0 | X_2 = u, Z = z] P[X_2 = u, Z = z]}{\sum_{u^*, z} P[Y_2 = 0 | Y_1 = 0, X_2 = u^*, Z = z] P[Y_1 = 0 | X_2 = u, Z = z] P[X_2 = u^*, Z = z]}
\end{aligned}
$$

We can see that the equality requires $P[Y_2 = 1 | Y_1 = 0, X_2 = u, Z = z] = 1 - P[Y_2 = 0 | Y_1 = 0, X_2 = u, Z = z]$ to be constant with respect to u and z, which is achieved when condition 12 is satisfied.

References

1. Falconer DS. Introduction to quantitative genetics. 3rd ed. LongmanWiley; 1989.

2. Papachristou C, Ober C, Abney M. Genetic variance components estimation for binary traits using multiple related individuals. Genet Epidemiol. 2011;35(5):291–302.

3. Bureau A, Croteau J, Couture C, Vohl MC, Bouchard C, Perusse L. Estimating genetic effect sizes under joint disease-endophenotype models in presence of gene-environment interactions. Front Genet. 2015;6:248.