Tradeoff Between No-Call Reduction in Genotyping Error Rate and Loss of Sample Size for Genetic Case/Control Association Studies

S.J. Kang, D. Gordon, A.M. Brown, J. Ott, and S.J. Finch

Pacific Symposium on Biocomputing 9:116-127(2004)

TRADEOFF BETWEEN NO-CALL REDUCTION IN GENOTYPING ERROR RATE AND LOSS OF SAMPLE SIZE FOR GENETIC CASE/CONTROL ASSOCIATION STUDIES

S. J. KANG[1], D. GORDON[2], A. M. BROWN[3], J. OTT[2], AND S. J. FINCH[1]
[1]*Department of Applied Mathematics and Statistics, State University of New York at Stony Brook, Stony Brook, NY 11794*
[2]*Laboratory of Statistical Genetics, Rockefeller University*
1230 York Avenue, New York, NY 10021-6399
[3]*Burke Medical Research Institute*
White Plains, NY 10605

Single nucleotide polymorphisms (SNP) may be genotyped for use in case-control designs to test for association between a SNP marker and a disease using a 2 × 3 chi-squared test of independence. Genotyping is often based on underlying continuous measurements, which are classified into genotypes. A "no-call" procedure is sometimes used in which borderline observations are not classified. This procedure has the simultaneous effect of reducing the genotype error rate and the expected number of genotypes observed. Both quantities affect the power of the statistic. We develop methods for calculating the genotype error rate, the expected number of genotypes observed, and the expected power of the resulting test as a function of the no-call procedure. We examine the statistical properties of the chi-squared test using a no-call procedure when the underlying continuous measure of genotype classification is a three-component mixture of univariate normal distributions under a range of parameter specifications. The genotype error rate decreases as the no-call region is increased. The expected number of observations genotyped also decreases. Our key finding is that the expected power of the chi-squared test is not sensitive to the no-call procedure. That is, the benefits of reduced genotype error rate are almost exactly balanced by the losses due to reduced genotype observations. For an underlying univariate normal mixture of genotype classification to be analyzed with a 2 × 3 chi-squared test, there is little, if any, increase in power using a no-call procedure.

1 Introduction

Single nucleotide polymorphisms (SNPs) genotypes are often determined by scoring technologies that first report the genotypes by one or more quantitative measurements[1,2]. Since the continuous measurements must be reduced to one of three genotypes (in this work, denoted by AA, AB, BB), some values may have ambiguous classification. One possible treatment of such classification is a "no-call" response; that is, no genotype is returned for the subject. For example, van den Oord et al.[3] comment that "technicians will not score points [genotypes] when the are segregated from the group". Throughout this work, we shall distinguish between the terms no-call and "all-call", where the latter indicates a procedure where all subjects are assigned a genotype, even if a subset are incorrect.

Some technologies classify genotypes using a mixture of univariate[2,4] or bivariate normal distributions[1,5]. For example, the Perkin Elmer software SNPscorer[5] uses an ellipsoidal model that they label "Ellipsoidal model of equal dimensions at constant orientation". This bivariate model could be reduced to modeling a mixture of univariate normal distributions by an appropriate projection. That is, the univariate data or the bi-variate data after projection follows the pattern shown in figure 1.

One standard use of SNP genotypes is a case/control genetic association analysis using the 2 × 3 chi-squared test of independence. We have previously investigated the effects of genotyping errors on the power of this test[6,7]. The major finding was that an increase in genotype error rates always resulted in a loss of power. The rationale for a no-call procedure is that the gain in power due to reduction of genotype error rates more than offsets the inevitable loss of power due to decrease in the number of genotype observations. In this work, we develop a method of computation to investigate this tradeoff. Specifically, we calculate both the genotype error rates and the reduction in expected sample size as a function of the no-call procedure. We then use these quantities to calculate the power of the test as a function of the no-call procedure.

2 Materials and Methods

2.1 Notation

The following notation is used through the remainder of this work:

Count variables:
☒ = Number of cases assuming no misclassification of genotypes (fixed)
N_U = Number of controls assuming no misclassification of genotypes (fixed)
N_A^{nc} = Number of cases adjusted after allowing no-call regions (random variable)
N_U^{nc} = Number of controls adjusted after allowing no-call regions (random variable)

Probability parameters:
p_a = Allele frequency of SNP marker *B* allele in the case group (affecteds)
p_u = Allele frequency of SNP marker *B* allele in the control group (unaffecteds)
P_{ij} = Frequency of SNP marker genotype j assuming no misclassification of genotypes ($i = 0$ for affected, $i = 1$ for unaffected, j=1 for *AA* genotype, j = 2 for *AB* genotype, j= 3 for *BB* genotype)
$P_{ij}^{'}$ = Probability of calling ij under no-call rule ($i = 0$ for affected, $i = 1$ for unaffected, j =1 for *AA* genotype, j = 2 for *AB* genotype, j = 3 for *BB* genotype), $P_{ij}^{'} < 1$

Expected proportion of subjects genotyped under the no-call rule:

$= P'_{01} + P'_{02} + P'_{03} < 1,$ with no-call rule.

$P'_{control} = P'_{11} + P'_{12} + P'_{13} < 1,$ with no-call rule.

$P^{nc}_{ij} =$ Probability of calling ij under no-call rule after adjustment ($i = 0$ for affected, $i = 1$ for unaffected, $j = 1$ for *AA* genotype, $j = 2$ for *AB* genotype, $j = 3$ for *BB* genotype)

These probabilities sum to one and are the genotype frequencies of each genotype conditional on the set of genotypes that are called.

Three component normal mixture parameters (see figure 1):

X = continuous measurement that is the underlying datum for classification into genotype

d_L = Mean of left-most (i.e., *AA*) genotype measurement, $d_L < 0$

d_R = Mean of right-most (i.e., *BB*) genotype measurement, $d_R > 0$

(Note: we set the mean of the heterozygote (i.e., *AB*) genotype measurement to be 0 and the variance of each component to 1)

= Half-width of the left no-call region

γ_R = Half-width of the right no-call region

c_L = Classification division point between *AA* and *AB* genotype with no-call; a subject is reported to have genotype *AA* when $X < c_L$ -

c_R = Classification division point between *AB* and *BB* genotype with no-call; a subject is reported to have genotype *BB* when $X > c_R + \gamma_R$

(Note: a subject is reported to have genotype *AB* when $c_L + \gamma_L < X < c_R - \gamma_R$)

$\Phi(\,)$ = Cumulative distribution function of standard normal random variable

$\phi(\,)$ = Probability density function of standard normal random variable

Error model functions:

The error rates are functions of the half-width parameters and γ_R.

$\varepsilon_{12}(\gamma_L, \gamma_R)$ = Pr (*AA* incorrectly coded as *AB* using no-call rule)

$\varepsilon_{13}(\gamma_L, \gamma_R)$ = Pr (*AA* incorrectly coded as *BB* using no-call rule)

$\varepsilon_{21}(\gamma_L, \gamma_R)$ = Pr (*AB* incorrectly coded as *AA* using no-call rule)

= Pr (*AB* incorrectly coded as *BB* using no-call rule)

$\varepsilon_{31}(\gamma_L, \gamma_R)$ = Pr (*BB* incorrectly coded as *AA* using no-call rule)

$\varepsilon_{32}(\gamma_L, \gamma_R)$ = Pr (*BB* incorrectly coded as *AB* using no-call rule)

Cost functions:

Similarly, the cost of each type of error is a function of the half-width parameters and γ_R.

$C_{ij}(\gamma_L,\gamma_R)$ = Cost of misclassifying i^{th} genotype in ordered list { AA, AB, BB}as j^{th} genotype in same list using no-call rule. For example:

$C_{12}(\gamma_L,\gamma_R)$ = Cost of misclassifying AA as AB using no-call rule

$f(\gamma_L,\gamma_R)$ = Fractional increase in sample size required to maintain asymptotic power.

$$f(\gamma_L,\gamma_R) = C_{12}(\gamma_L,\gamma_R)\varepsilon_{12}(\gamma_L,\gamma_R) + C_{13}(\gamma_L,\gamma_R)\varepsilon_{13}(\gamma_L,\gamma_R) + C_{21}(\gamma_L,\gamma_R)\varepsilon_{21}(\gamma_L,\gamma_R)$$
$$+ C_{23}(\gamma_L,\gamma_R)\varepsilon_{23}(\gamma_L,\gamma_R) + C_{31}(\gamma_L,\gamma_R)\varepsilon_{31}(\gamma_L,\gamma_R) + C_{32}(\gamma_L,\gamma_R)\varepsilon_{32}(\gamma_L,\gamma_R)$$

K = Random variable that is ratio of controls to cases after no-call rule.

(Note: is a fixed aspect of design, $K = \dfrac{N_A^{nc}}{N_U^{nc}}$ is a random variable with mean approximately equal to k, N_A^{nc} is a binomial random variable on N_A trials with probability of call , and similarly N_U^{nc} is a binomial random variable.)

2.2 *Computation of genotype error rates and genotype frequencies with no-call region*

The error rates for the genotype classification with no-call regions assuming a three-component univariate normal mixture are:

$$\varepsilon_{12}(\gamma_L,\gamma_R) = \Phi(c_R - d_L - \gamma_R) - \Phi(c_L - d_L + \gamma_L)$$
$$\varepsilon_{21}(\gamma_L,\gamma_R) = \Phi(c_L - \gamma_L)$$
$$\varepsilon_{23}(\gamma_L,\gamma_R) = 1 - \Phi(c_R + \gamma_R)$$
$$\varepsilon_{32}(\gamma_L,\gamma_R) = \Phi(c_R - d_R - \gamma_R) - \Phi(c_L - d_R + \gamma_L)$$
$$\varepsilon_{13}(\gamma_L,\gamma_R) =$$
$$\varepsilon_{31}(\gamma_L,\gamma_R) = \Phi(c_L - d_R - \gamma_L)$$

In the results section, we present computations of the error rates as functions of the no-call region half-widths γ_L and γ_R (see discussion of figure 2 in Results). Note as mentioned above that the error rates are functions of the means and cut-points.

2. 3 *Probability of calling each genotype in the presence of errors with no-call half-width*

$$P_{01}^{'} = P_{01}\Phi(c_L - \gamma_L - d_L) + P_{02}\Phi(c_L - \gamma_L) + P_{03}\Phi(c_L - \gamma_L - d_R)$$

$$P'_{02} = P_{01}\{\Phi(c_R-\gamma_R-d_L)-\Phi(c_L+\gamma_L-d_L)\} + P_{02}\{\Phi(c_R-\gamma_R)-\Phi(c_L+\gamma_L)\} + P_{03}\{\Phi(c_R-\gamma_R-d_R)-\Phi(c_L+\gamma_L-d_R)\}$$

$$= P_{01}\{1-\Phi(c_R+\gamma_R-d_L)\} + P_{02}\{1-\Phi(c_R+\gamma_R)\} + P_{03}\{1-\Phi(c_R+\gamma_R-d_R)\}$$

$$P'_{11} = P_{11}\Phi(c_L-\gamma_L-d_L)+P_{12}\Phi(c_L-\gamma_L)+P_{13}\Phi(c_L-\gamma_L-d_R)$$

$$P'_{12} = P_{11}\{\Phi(c_R-\gamma_R-d_L)-\Phi(c_L+\gamma_L-d_L)\} + P_{12}\{\Phi(c_R-\gamma_R)-\Phi(c_L+\gamma_L)\} + P_{13}\{\Phi(c_R-\gamma_R-d_R)-\Phi(c_L+\gamma_L-d_R)\}$$

$$P'_{13} = P_{11}\{1-\Phi(c_R+\gamma_R-d_L)\} + P_{12}\{1-\Phi(c_R+\gamma_R)\} + P_{13}\{1-\Phi(c_R+\gamma_R-d_R)\}$$

We rescale each probability by and $P'_{control}$ so that $P^{nc}_{01}+P^{nc}_{02}+P^{nc}_{03}=1$ and . That is,

$P^{nc}_{01} = \frac{P'_{01}}{P'_{case}}$, $P^{nc}_{11} = \frac{P'_{11}}{P'_{control}}$, $P^{nc}_{02} = \frac{P'_{02}}{P'_{case}}$, and so forth. Note that these probabilities are conditional probabilities.

2.2 *Non-centrality parameter*

The non-centrality parameter for the all-call situation is

$$\lambda_1 = N_A N_U\{\frac{(P_{01}-P_{11})^2}{N_A P_{01}+N_U P_{11}} + \frac{(P_{02}-P_{12})^2}{N_A P_{02}+N_U P_{12}} + \frac{(P_{03}-P_{13})^2}{N_A P_{03}+N_U P_{13}}\}$$

The non-centrality parameter for no-call, $\lambda_2(N^{nc}_A, N^{nc}_U)$, is a random variable; that is, it is a function of the random variables N^{nc}_A and N^{nc}_U given by:

$$N^{nc}_A N^{nc}_U\{\frac{(P^{nc}_{01}-P^{nc}_{11})^2}{N^{nc}_A P^{nc}_{01}+N^{nc}_U P^{nc}_{11}} + \frac{(P^{nc}_{02}-P^{nc}_{12})^2}{N^{nc}_A P^{nc}_{02}+N^{nc}_U P^{nc}_{12}} + \frac{(P^{nc}_{03}-P^{nc}_{13})^2}{N^{nc}_A P^{nc}_{03}+N^{nc}_U P^{nc}_{13}}\}$$

Using the "delta method"[8], $E(\lambda_2(N^{nc}_A, N^{nc}_U))$ is approximately equal to:

$$\frac{N_A N_U}{P'_{case}P'_{control}}\{\frac{(P'_{control}P'_{01}-P'_{case}P'_{11})^2}{N_A P'_{01}+N_U P'_{11}} + \frac{(P'_{\text{cot rol}}P'_{02}-P'_{case}P'_{12})^2}{N_A P'_{02}+N_U P'_{12}} + \frac{(P'_{control}P'_{03}-P'_{case}P'_{13})^2}{N_A P'_{03}+N_U P'_{13}}\}$$

We define *expected power* to be the power of the chi-squared test using the expected non-centrality parameter.

2.3 *Cost function with symmetric no-call regions*

When , the cost function $f(\gamma_L, \gamma_R) = f(\gamma)$ is:

$$C_{12}(\gamma)\varepsilon_{12}(\gamma)+C_{13}(\gamma)\varepsilon_{13}(\gamma)+C_{21}(\gamma)\varepsilon_{21}(\gamma)+C_{23}(\gamma)\varepsilon_{23}(\gamma)+C_{31}(\gamma)\varepsilon_{31}(\gamma)+C_{32}(\gamma)\varepsilon_{32}(\gamma)$$

$$= C_{12}\{\Phi(c_R-d_L-\gamma)-\Phi(c_L-d_L+\gamma)\}+C_{21}\Phi(c_L-\gamma)+C_{13}\{1-\Phi(c_R-d_L+\gamma)\}$$

$$+C_{31}\Phi(c_L-d_R-\gamma)+C_{23}\{1-\Phi(c_R+\gamma)\}+C_{32}\{\Phi(c_R-d_R-\gamma)-\Phi(c_L-d_R+\gamma)\},$$

where

$$C_{12}(\gamma_L,\gamma_R)=\frac{1}{g}\times\frac{[(P_{01}^{nc}-P_{11}^{nc})(P_{02}^{nc}+KP_{12}^{nc})-(P_{02}^{nc}-P_{12}^{nc})(P_{01}^{nc}+KP_{11}^{nc})]^2}{(P_{01}^{nc}+KP_{11}^{nc})(P_{02}^{nc}+KP_{12}^{nc})^2},$$

$$C_{13}(\gamma_L,\gamma_R)=\frac{1}{g}\times\frac{[(P_{01}^{nc}-P_{11}^{nc})(P_{03}^{nc}+KP_{13}^{nc})-(P_{03}^{nc}-P_{13}^{nc})(P_{01}^{nc}+KP_{11}^{nc})]^2}{(P_{01}^{nc}+KP_{11}^{nc})(P_{03}^{nc}+KP_{13}^{nc})^2},$$

$$C_{21}(\gamma_L,\gamma_R)=\frac{1}{g}\times\frac{[(P_{02}^{nc}-P_{12}^{nc})(P_{01}^{nc}+KP_{11}^{nc})-(P_{01}^{nc}-P_{11}^{nc})(P_{02}^{nc}+KP_{12}^{nc})]^2}{(P_{02}^{nc}+KP_{12}^{nc})(P_{01}^{nc}+KP_{11}^{nc})^2},$$

,

$$C_{31}(\gamma_L,\gamma_R)=\frac{1}{g}\times\frac{[(P_{03}^{nc}-P_{13}^{nc})(P_{01}^{nc}+KP_{11}^{nc})-(P_{01}^{nc}-P_{11}^{nc})(P_{03}^{nc}+KP_{13}^{nc})]^2}{(P_{03}^{nc}+KP_{13}^{nc})(P_{01}^{nc}+KP_{11}^{nc})^2},$$

$$C_{32}(\gamma_L,\gamma_R)=\frac{1}{g}\times\frac{[(P_{03}^{nc}-P_{13}^{nc})(P_{02}^{nc}+KP_{12}^{nc})-(P_{02}^{nc}-P_{12}^{nc})(P_{03}^{nc}+KP_{13}^{nc})]^2}{(P_{03}^{nc}+KP_{13}^{nc})(P_{02}^{nc}+KP_{12}^{nc})^2},$$

and $g=\{\frac{(P_{01}^{nc}-P_{11}^{nc})^2}{P_{01}^{nc}+KP_{11}^{nc}}+\frac{(P_{02}^{nc}-P_{12}^{nc})^2}{P_{02}^{nc}+KP_{12}^{nc}}+\frac{(P_{03}^{nc}-P_{13}^{nc})^2}{P_{03}^{nc}+KP_{13}^{nc}}\}$.

Research Approach

We evaluate three functions: (1) the error rates for the genotype classification with no-call regions (figure 2); (2) the expected proportion of subjects genotyped, and $P_{control}^{'}$, under the no-call procedure (also figure 2); and (3) the expected power under the no-call procedure (figure 3). Note that all three are functions of the no-call region half-widths γ_L and γ_R (see discussion of figure 2 in Results).

Our main question is: is there a setting of the no-call region half-widths γ_L and γ_R that maximize the power of the chi-squared test? We investigate this question for the parameter settings $N_A=N_U=500$, $P_a=0.2, P_u=0.15$ with both groups in Hardy Weinberg equilibrium, $d_R=-d_L=2$, , $c_L=\frac{d_L}{2}$, and level of significance 0.05. We note that, while we assume symmetric means and cut-points for our examples, our methodology is completely general and may be applied to non-symmetric means and cut-points as well.

3 Results

The error rates decrease as the half-width of the no-call regions increase. Figure 2 presents the error rates for the genotype classification with no-call regions and the expected proportion of subjects genotyped, and $P'_{control}$, under the no-call rule, when the means are symmetric (i.e., $d_R = -d_L$=2), the cut points are half way between the means (i.e., , $c_L = \frac{d_L}{2}$) and the no-call half-widths are equal (i.e, $\gamma_L = \gamma_R = \gamma$). Note that under these conditions $\varepsilon_{12} = \varepsilon_{32}$ and $\varepsilon_{21} =$. This figure documents that genotype error rates decrease as the no-call half-width γ increases, and that simultaneously, the expected proportion of subjects genotyped under the no-call rule decreases. In the extreme case, when the half-width of the no-call regions is indefinitely large, the probabilities of misclassification are zero, but no observations are genotyped. The power of the 2 x 3 chi-squared test on genotypes generated using a no-call rule with indefinitely large half-widths is zero. Thus, we search for a setting of γ that maximizes the expected power of the chi-squared test.

In figure 3, we plot the expected power of the chi-squared test using the parameter settings described above (Methods – Research Approach) to understand the trade-off between lowered probability of misclassification using larger γ's and the lowered sample size due to the decrease in the number of observations. In general, the power is not sensitive to the choice of half-width. That is, the gain from the reduction in the probabilities of misclassification is almost exactly balanced by the loss in the number of observations genotyped for the 2 x 3 chi-squared test. In some situations, there is a small gain in power due to using an optimal choice of half-width. For example, setting γ_L=0.25 and γ_R= 0.0 for the scenario in figure 3 returns an expected power of 0.481, an increase of 0.02 over the all-call rule. A choice of half-width other than the optimal half-widths for a situation is associated with a small loss of power.

While we do not report the findings here, we note that we performed the above-mentioned analyses for a range of settings for the parameters γ_L, γ_R, d_R, d_L, P_a, P_u and *k*. Our results for different parameter settings were essentially the same as the ones presented here (data not shown).

Discussion

Use of a no-call rule lowers the probability of misclassification but also lowers the number of observations classified. With respect to the expected power of the 2×3 chi-squared test of equality of genotype frequencies in cases and controls, the gain in power from the reduction in the probabilities of misclassification almost exactly matches the loss in power due to the reduction in the number of observations used for the parameter settings we investigated. The use of Occam's razor suggests that a no-call rule not be used when the data generated are analyzed with the 2×3 chi-squared test. That is, any gain in power from using a no-call rule will be small compared to the power of the chi-squared test based on using all observations, and there may well be a net loss in power due to reduction in sample size. The extra effort in using a no-call procedure produces no clear-cut gain in power when the chi-squared chi-squared test is used.

We hypothesize that a similar finding holds for other tests, such as Armitage's test for trend in proportions[9]. The question can be answered by following the steps in this analysis.

The results presented in this paper use situations with considerable symmetry. For example, in figure 2, component means were symmetric ($d_R = -d_L$), cut-points were symmetric (, $c_L = \frac{d_L}{2}$), and half-widths of the no-call regions were equal ($\gamma_L = \gamma_R = \gamma$). This setting was made to reduce the dimensionality of the graphics so that they focused on the essence of the findings. In figure 3, for example, the half-widths of the left and right no-call regions ranged over all possible values in the specified ranges. Space limitations preclude presenting a comprehensive set of graphs. These graphs are available on our website (http://linkage.rockefeller.edu/derek/psb2004.html).

Also, we note that SNP genotype assignments can be modeled with by a mixture of three bi-variate normal distributions[1]. In such a model, the probability of misclassifying one homozygote as another may be non-negligible. Since such errors are much more costly[10], there may well be situations in which use of a no-call rule is in fact advantageous. The techniques in this paper can be readily applied to such a situation, and we are currently performing such calculations. Another issue regarding the transformation from a mixture of bi-variate distributions to a mixture of univariate distributions is that the order of genotypes may be different from the order presented in figure 1 (e.g., the two homozygote distributions may be adjacent). However, for the references that we have investigated[1,5], a projection onto a line $Ax + By = C$, where A, $B > 0$, will always result in an ordering of genotypes as we have presented in figure 1.

Acknowledgments

The authors gratefully acknowledge grants K01-HG00055 and R01-MH59492 from the National Institutes of Health.

References

1. Ranade, K. et al. High-throughput genotyping with single nucleotide polymorphisms. *Genome Research* **11**, 1262-8 (2001).
2. Ahmadian, A., Gharizadeh, B., O'Meara, D., Odeberg, J. & Lundeberg, J. Genotyping by apyrase-mediated allele-specific extension. *Nucleic Acids Res* **29**, E121 (2001).
3. van den Oord, E.J.C.G., Jiang, Y., Riley, B.P., Kendler, K.S. & Chen, X. FP-TDI SNP scoring by manual and statistical procedures: A study of error rates and types. *Biotechniques* **34**, 610-6, 618-20, 622 (2003).
4. O'Meara, D., Ahmadian, A., Odeberg, J. & Lundeberg, J. SNP typing by apyrase-mediated allele-specific primer extension on DNA microarrays. *Nucleic Acids Res* **30**, e75 (2002).
5. Grant, S.F. et al. SNP genotyping on a genome-wide amplified DOP-PCR template. *Nucleic Acids Res* **30**, e125 (2002).
6. Gordon, D., Finch, S.J., Nothnagel, M. & Ott, J. Power and sample size calculations for case-control genetic association tests when errors are present: application to single nucleotide polymorphisms. *Human Heredity* **54**, 22-33 (2002).
7. Gordon, D., Levenstien, M.A., Finch, S.J. & Ott, J. Errors and linkage disequilibrium interact multiplicatively when computing sample sizes for genetic case-control association studies. in *Pacific Symposium on Biocomputing* 490-501 (2003).
8. Casella, G. & Berger, R.L. *Statistical Inference*, (Duxbury-Thomson Learning, Pacific Grove, CA, 2002).
9. Armitage, P. Tests for linear trends in proportions and frequencies. *Biometrics* **11**, 375-386 (1955).
10. Kang, S.J., Gordon, D. & Finch, S.J. What SNP genotyping errors are most costly for genetic association studies? *Genetic Epidemiology* **(in press)**(2003).