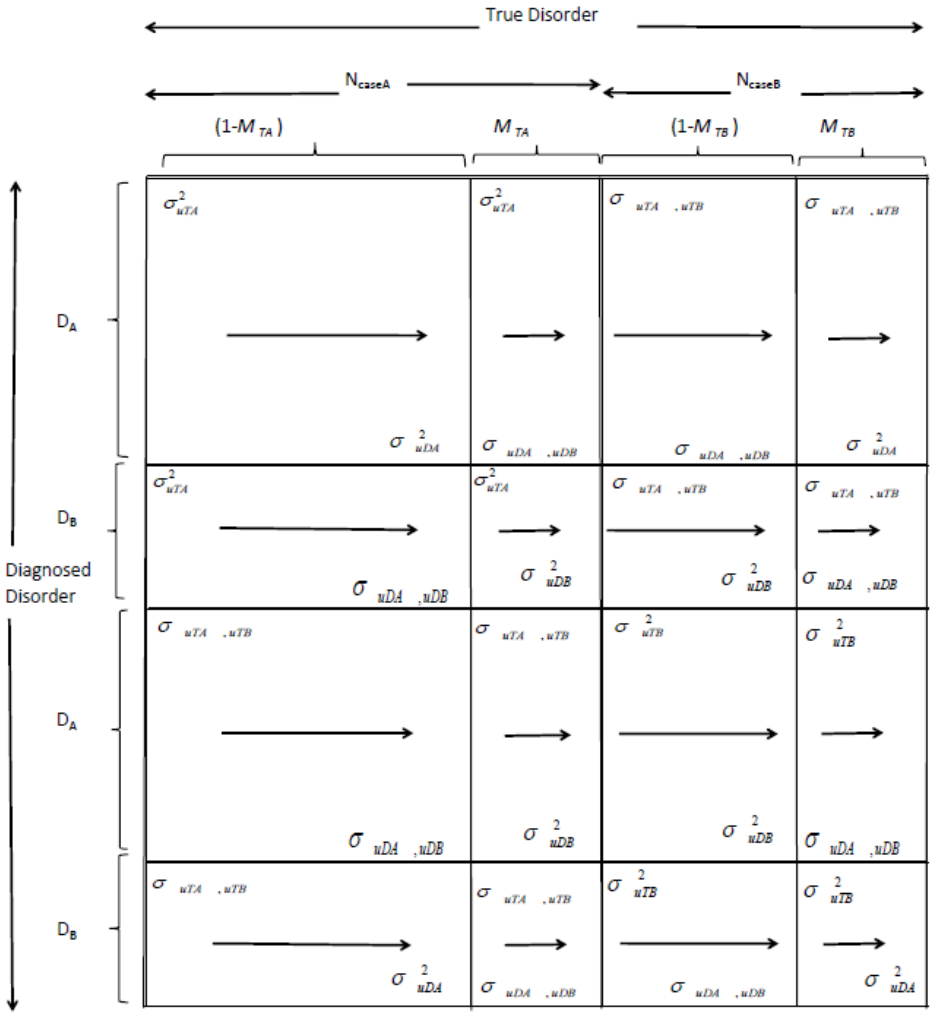
Supplementary Information

Impact of diagnostic misclassification on estimation of genetic correlations using genome-wide genotypes

Naomi R Wray[1,2*], S Hong Lee[1], Kenneth S Kendler[3]

1. Queensland Institute of Medical Research, Herston Road, Brisbane 4029, Australia
2. Queensland Brain Institute, University of Queensland, St Lucia, 4067, Australia
3. Department of Psychiatry, Medical College of Virginia/Virginia Commonwealth University, Richmond, Virginia, USA

Supplementary Figure 1. Schematic representation of the case-case diagonal block of the **A** matrix which is a square matrix with dimensions $N_{caseA} + N_{caseB}$. The elements of this matrix are the genome-wide genetic similarities between pairs of cases. The schematic shows the elements that contribute to the estimation of the variance of the true disorders and the elements that contribute to the estimation of the variance of the diagnosed disorders. The numbers of elements in each block are the weights given to the true disorder variances/covariances as allocated to the diagnosed disorder variance/covariances.



Impact of misdiagnosis on power of genome-wide association studies.

Our main interest has been the impact of misdiagnosis on estimation of genetic parameters either from family data or from genome-wide genotypes. Here we consider the impact of misdiagnosis on the power of detection of individual risk loci in genetic association analysis. As before we have a disorder which has lifetime probability K_T . We consider a causal variant with frequency of the risk allele and protective alleles of p and (1-p) respectively in the population. Let $(1 - p)^2$, $2p(1 - p)$ and p^2 be the frequencies of the genotypes (in Hardy-Weinberg equilibrium), with risks of f_0, f_1 and f_2. If we assume a multiplicative model on this disease scale, then $f_1 = f_0 \gamma$ and $f_2 = f_0 \gamma^2$ where γ is the relative risk of the risk allele compared to the protective allele. We can calculate the frequency of the risk alleles in cases (true cases) and screened controls as

and $\frac{p}{1-K}(1-\frac{K\gamma}{1+p(\gamma-1)})$

$p_{caseT} = \frac{p\gamma}{1+p(\gamma-1)}$ and $p_{control} = \frac{p}{1-K}(1-\frac{K\gamma}{1+p(\gamma-1)})$.

If M_D is the proportion of cases that are misdiagnosed then

$p_{caseD} = (1- M_D)\, p_{caseT} + M_D p_{control}$

The non-centrality parameter (NCP) of the X^2 test of association is

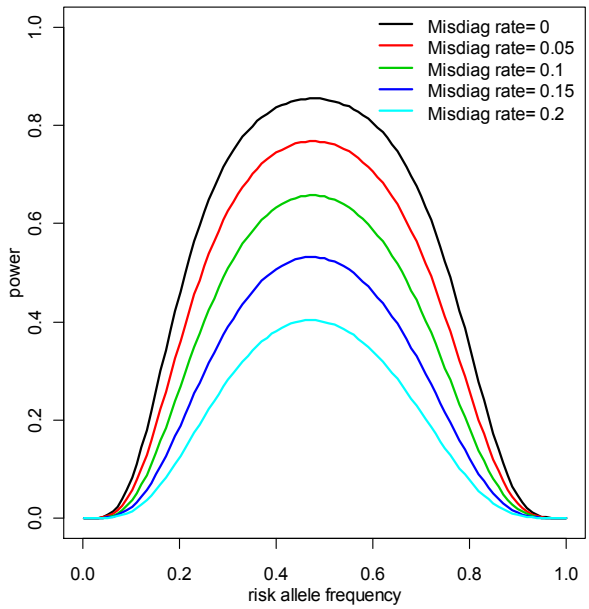
$$\text{NCP} = \frac{N^2(p_{caseD} - p_{control})^2}{Var(\hat{p}_{caseD} - \hat{p}_{control})} = \frac{Nv(1-v)(p_{caseD} - p_{control})^2}{\bar{p}_D(1-\bar{p}_D)}$$

where $\bar{p} = v\, p_{caseD} + (1-v)\, p_{control}$ and where $v = N_{case}/(N_{case} + N_{control}) = N_{case}/N$ and p is the allele frequency of the allele in the sample denoted by the subscript, i.e. p_{caseD} is the allele frequency in the sample diagnosed as cases. We calculate power as the normal probability p(Z > T), where $Z = \sqrt{NCP}$ and T is the normal deviate corresponding to the type I probability

level, i.e., 5 x 10^{-8} for genome-wide association. When $M_D = 0$, the power calculation agrees with the genetic power calculator [1].

We quantify the impact of misclassification of disorders on the power of genome-wide association studies. A case-control study of 5000 diagnosed cases and 5000 controls of a disorder with true lifetime risk of 1% has ~ 84% power to detect a risk variant of frequency 0.4 and relative risk 1.2 at the genome-wide significance threshold of $p< 5x10^{-8}$, but the power reduces to only 64% or ~38% when 10% or 20% of cases, respectively, have been misdiagnosed (Supplementary Figure 2).

Supplementary Figure 2. Impact of misdiagnosis on the power of genetic association studies, assuming 5000 diagnosed cases and 5000 controls for a disease with lifetime risk of 1%, a genotype relative risks of 1.2 and 1.2^2 for the heterozygote and risk allele homozygote, respectively, and a type I error threshold of $5x10^{-8}$.



REFERENCES

1. Purcell S, Cherny SS, Sham PC: Genetic Power Calculator: design of linkage and association genetic mapping studies of complex traits. *Bioinformatics* 2003; **19:** 149-150.