Genetic epidemiological approaches in complex neurological disorders

Hottenga, J.J.

CHAPTER 3

A straightforward approach to overcome false-positive associations in studies of gene-gene interaction

JJ Hottenga[1,2], G Roks[1], B Dermaut[3], C van Broeckhoven[3], CM van Duijn[1,3]

1. Department of Epidemiology & Biostatistics, Erasmus Medical Center Rotterdam, the Netherlands.
2. Department of Human Genetics, Leiden University Medical Center, Leiden, the Netherlands.
3. Department of Molecular Genetics, Flanders Interuniversity Institute for Biotechnology, University of Antwerp, Antwerp, Belgium.

For submission

Abstract

Research of gene-gene interactions will be important in unraveling genetic risk factors involved in complex traits. However, based on association studies, gene interactions are susceptible to false-positive as well as false-negative findings. One of the reasons may be stratification of a limited number of cases and controls, leading to a small number of subjects in each stratum. In particular the number of controls who carry the risk allele of both genes studied is often small, even for common polymorphisms. We present a straightforward approach to reveal spurious associations due to over-stratification that can be applied in gene-gene interaction studies. Basically, we propose to analyze cases and controls separately. In controls, association between two unlinked genes will indicate bias in the findings of the study. From this approach it also follows that one may improve the statistical power and reduce the probability of false-positive findings by genotyping controls for the second gene in the limiting stratum specifically, i.e. control carriers of the risk allele of the first gene studied. This approach may be useful in large-scale epidemiological studies, in which multiple genes often have been characterized. We will illustrate the approach using an example of a study of interaction of the Apolipoprotein E and Presenilin-1 gene in relation to Alzheimer's disease. In this study, a false-positive association was detected using this method.

Keywords

Gene-gene interaction, Apolipoproteins E, Presenilin-1, Alzheimer's disease, interaction test.

Introduction

Candidate gene research has shown to be prone to false-positive findings, despite its potential for studies of complex genetic disorders[1,2]. Typically these studies target polymorphisms in genes, which in effect may depend, for a large part, on other genes. Studies of gene-gene interactions are therefore common practice in genetic epidemiological research. To this end, the effect of a polymorphism is often studied in a stratified analysis based on the presence of a second allele of a polymorphism. In this paper we argue that as a result of the stratification, false-positive findings may occur because of too small numbers in the stratified groups. A simple straightforward approach is presented to evaluate this form of bias. Furthermore, the statistical power when using this approach was examined and the increase shows that it may also prevent false-negative studies. To illustrate the method, a study of interaction between the Apolipoprotein E (APOE) and Presenilin-1 (PSEN1) genes in Alzheimer's disease (AD) is presented. The APOE ε4 allele (APOE*4) has shown to be an established risk factor for AD[3]. The PSEN1 gene can be mutated in patients with early-onset familial AD[4]. Findings on polymorphisms in the PSEN1 gene in relation to late-onset sporadic AD have been inconsistent[5–8]. In this study an interaction was found between APOE*4, PSEN1 and AD. However, when evaluated with the approach, the result was shown to be a false-positive.

Methods

Approach

In studies of gene-gene interaction, polymorphisms of two genes are studied for a specific disorder. Data of the two polymorphisms can be analyzed under several interaction models[9,10]. Table 1A illustrates an often used approach, in which the data are stratified for the presence of allele *X* for gene 1 and allele *Y* for gene 2. If there is interaction present between the genes, the odds ratio for gene 2 will be different for the strata of gene 1. These odds ratios are given as (*A x D*) / (*B x C*) and (*E x H*) / (*G x F*). If the studied genes are unlinked, for

example because they are located on different chromosomes, then according to Mendelian laws the two genes will segregate independently from each other. For most combinations of genes that are studied in interaction studies this is true. Consequently, the alleles of the two genes, when rewriting table 1A into 1B stratifying by case-control status, should not show association in controls (table 1B)[11,12]. A significant deviation of the odds ratio from one in controls is not compatible with Mendelian segregation of alleles. There may be some explanations, for example mixture of genetically different populations. However, a more likely explanation to consider is that the finding is a false-positive result due to the stratification of the controls into small subgroups. *A priori* one expects to find a low number of controls in the individual strata, especially in the cell having the two risk alleles associated with the studied disorder (cell *G* in tables 1A and 1B). This problem can be overcome by increasing the number of controls in the study[13,14]. Controls can be added at random enlarging the total number of controls. However, in large-scale epidemiological studies in which several genes have already been examined, controls can also be added to specific strata for which there are specifically unstable low numbers. This is illustrated in table 1C, in which the control series in the stratum of allele *X* carriers of gene 1 is increased by a factor *K*, yielding an odds ratio of $(C \times KH) / (D \times KG) = (C \times H) / (D \times G)$ similar to the one found in table 1B. Therefore, with either method of adding controls the new odds ratio will remain unbiased, with the second approach being a more cost effective alternative. The use of this approach was illustrated in an example of AD.

Example

In this example, the empirical data comprise a series of patients with late-onset Alzheimer's disease, which were drawn from the Rotterdam study[15]. This is a population-based prospective study of over 8000 residents aged 55 years and older of a suburb in Rotterdam in the Netherlands. All participants in the study gave informed consent.

Table 1

Odds ratios in a case- control study of gene-gene interaction.

A

Gene 2		Gene 1 allele *X*+	allele *X*-	Odds ratio
allele *Y*-	cases	*A*	*B*	*(A x D) / (B x C)*
	controls	*C*	*D*	
allele *Y*+	cases	*E*	*F*	*(E x H) / (F x G)*
	controls	*G*	*H*	

B

	Gene 2	Gene 1 allele *X*+	allele *X*-	Odds ratio
cases	allele *Y*-	*A*	*B*	*(A x F) / (B x E)*
	allele *Y*+	*E*	*F*	
controls	allele *Y*-	*C*	*D*	*(C x H) / (D x G)*
	allele *Y*+	*G*	*H*	

C

	Gene 2	Gene 1 allele *X*+	allele *X*-	Odds ratio
cases	allele *Y*-	*A*	*B*	*(A x F) / (B x E)*
	allele *Y*+	*E*	*F*	
controls	allele *Y*-	*C*	*D*	*(C x KH) / (D x KG)*
	allele *Y*+	*KG*	*KH*	

Table 1A is stratified for the presence of allele *X*, then for cases and controls.
Table 1B is stratified for cases and controls, then for the presence of allele *X*.
Table 1C the number of controls in the allele *X*+ stratum is increased by factor *K*.

The –22 C/T PSEN1 polymorphism located in the promoter region of the gene was genotyped in a nested case- control study including 316 AD patients and 219 age-matched controls[16,17]. Because APOE is a pivotal gene involved in lipid metabolism and neurodegenerative disorders, all of the participants of the Rotterdam study have been genotyped for the APOE gene previously for other studies[18]. No linkage disequilibrium is expected between the two genes, since the PSEN1 gene is located on chromosome 14q24.2 and the APOE gene on chromosome 19q13.31 [4,19]. The number of controls carrying APOE*4 was subsequently increased by a factor K = 1.96 by genotyping 54 additional

controls for the PSEN1 gene following our approach. Odds ratios were calculated for the strata of the APOE*4 allele and PSEN1 genotypes and interactions were tested using binary logistic regression. For the statistical analysis, CT and TT genotypes of the PSEN1 gene were pooled because of the low number of persons carrying the TT genotype[20]. All statistics were tested with SPSS for Windows 10.0.

Power

Power calculations were performed to illustrate the effects of adding controls in a single stratum. Hence, the number of controls with genetic risk factor *X* was increased for values of *K* ranging from 1 to 4 (table 1C). Increasing *K* more than four times the number of controls usually does not increase the power further[13,14]. The power to detect association for allele *Y* given *X* was calculated for 250 and 500 cases and controls, respectively. Allele frequencies of *X* and *Y* were varied with values of 0.10, 0.25 and 0.50. The relative risk of the disorder given *X* was assumed to be 2.00 and the additional disorder risk for *Y* given presence of *X* was 2.00 as well. The significance level α was set to 0.05 and the prevalence of the disorder was 0.10. No genetic model was presumed for both polymorphisms and analysis was assumed to be done with allelic tests of association. The power was calculated with the web-based program PAWE 1.2[21,22].

Results

Gene-gene interaction was studied between the APOE ε4 allele and the –22 C/T PSEN1 promoter polymorphism in relation to Alzheimer's disease. For both genes, no significant deviations from Hardy-Weinberg equilibrium were observed in cases and controls. Table 2A shows the association between the PSEN1 genotype and AD, stratified by the presence of the APOE*4 allele. The results show that the number of CC carriers in AD patients is reduced as compared to the controls in the stratum of APOE*4 carriers. The odds ratio

(OR) for carriers of the CT/TT genotype equaled 0.24 with a 95% confidence interval (CI) of 0.07 to 0.84. In subjects who did not carry the APOE*4 there was no association between PSEN1 and AD. The OR, pooling the CT and TT genotype, equals 1.14 (95% CI = 0.66-1.98). Testing for interaction between PSEN1 and APOE showed evidence for interaction using a multiplicative model (p = 0.025).

Table 2

A gene-gene interaction study of the Apolipoprotein E ε4 allele (APOE*4) and the –22 C/T Presenilin-1 promoter polymorphism in relation to Alzheimer's disease (AD).

A						
APOE		PSEN1 CC	CT+TT	Number	Odds ratio	95% CI
APOE*4-	AD	169 (84%)	32 (16%)	201	1.14	0.66-1.98
	controls	134 (82%)	29 (18%)	163		
APOE*4+	AD	93 (81%)	22 (19%)	115	0.24	0.07-0.84
	controls	53 (95%)	3 (5%)	56		

B						
	APOE	PSEN1 CC	CT+TT	Number	Odds ratio	95% CI
AD	APOE*4-	169 (84%)	32 (16%)	201	1.25	0.69-2.27
	APOE*4+	93 (81%)	22 (19%)	115		
Controls	APOE*4-	134 (82%)	29 (18%)	163	0.26	0.08-0.90
	APOE*4+	53 (95%)	3 (5%)	56		

APOE*4+ and APOE*4- = presence and absence of the Apolipoprotein E ε4 allele,
PSEN1 = the Presenilin-1 genotype, 95% CI = The 95% confidence interval of the odds ratio.
Table 2A is stratified for the presence of the APOE*4 allele, then for cases and controls.
Table 2B is stratified for cases and controls, then for the presence of the APOE*4 allele.

When examining the association between APOE and PSEN1 for cases and controls separately, evidence for association between the two genes was found only in the control group (OR = 0.26; 95% CI = 0.08-0.90) (table 2B). No evidence for association was found in cases (OR = 1.25; 95% CI = 0.69-2.27). As the APOE and PSEN1 genes are located on different chromosomes, the

association in controls is not compatible with Mendelian segregation. Therefore, a true interaction effect of the two genes is unlikely. In addition, differential survival related to the genotypes is unlikely, because it is assumed to occur both in cases and age-matched controls. The interaction observed in table 2A is most likely the result of the low number of controls carrying both the APOE*4 allele and the T allele at the PSEN1 promoter.

Table 3

A reanalysis of the gene-gene interaction study of the Apolipoprotein E ε4 allele (APOE*4) and the –22 C/T Presenilin-1 promoter polymorphism in relation to Alzheimer's disease (AD) with added controls.

		PSEN1				
	APOE	CC	CT+TT	Number	Odds ratio	95% CI
AD	APOE*4-	169 (84%)	32 (16%)	201	1.25	0.69-2.27
	APOE*4+	93 (81%)	22 (19%)	115		
Controls	APOE*4-	134 (82%)	29 (18%)	163	0.62	0.31-1.25
	APOE*4+	97 (88%)	13 (12%)	110		

APOE*4+ and APOE*4- = presence and absence of the Apolipoprotein E ε4 allele,
PSEN1 = the Presenilin-1 genotype. 95% CI = The 95% confidence interval of the odds ratio.

As all participants of the Rotterdam study were already genotyped for the APOE gene, we genotyped 54 extra controls for PSEN1 from the non-demented subjects carrying the APOE*4 allele, increasing the sample with factor $K = 1.96$. When adding these controls to the initial set, the frequency of the CC genotype decreased from 95% to 88%, while the number CT genotype carriers increased from 5% to 12% (table 3). The odds ratio analyzed in controls only, changed from 0.26 to 0.62 with the 95% confidence interval ranging from 0.31 to 1.25. This suggests that adding the controls overcame the problem presented in table 2B (table 3). As a result the PSEN1 genotype frequencies did not longer differ between cases and controls in the APOE*4 carrier stratum (OR = 0.57; 95% CI = 0.27-1.19). Testing for interaction using

logistic regression analysis showed no more evidence for interaction (p = 0.137).

Table 4

Power calculations for increasing the number of controls in a specific stratum of a gene-gene interaction case-control study.

F(*X*+)	N cases / controls[a]	N controls Random[b]	*K*	N cases *X*+[c]	N Controls *X*+[d]	Power F(*Y*+) = 0.10	Power F(*Y*+) = 0.25	Power F(*Y*+) = 0.50
0.10	250 / 250	250	1	45	23	0.286	0.507	0.527
	250 / 273	500	2	45	46	0.428	0.695	0.714
	250 / 296	750	3	45	69	0.518	0.784	0.785
	250 / 319	1250	4	45	92	0.583	0.816	0.823
	500 / 500	500	1	91	45	0.498	0.793	0.814
	500 / 545	1000	2	91	90	0.707	0.936	0.946
	500 / 590	1500	3	91	135	0.807	0.972	0.973
	500 / 635	2500	4	91	180	0.864	0.981	0.983
0.25	250 / 250	250	1	100	58	0.586	0.869	0.881
	250 / 308	500	2	100	116	0.788	0.968	0.972
	250 / 366	750	3	100	174	0.871	0.988	0.988
	250 / 424	1250	4	100	232	0.914	0.992	0.992
	500 / 500	500	1	200	117	0.871	0.992	0.994
	500 / 617	1000	2	200	234	0.974	1.000	1.000
	500 / 734	1500	3	200	351	0.992	1.000	1.000
	500 / 851	2500	4	200	468	0.997	1.000	1.000
0.50	250 / 250	250	1	167	120	0.861	0.990	0.991
	250 / 370	500	2	167	240	0.967	0.999	0.999
	250 / 490	750	3	167	360	0.988	1.000	1.000
	250 / 610	1250	4	167	480	0.994	1.000	1.000
	500 / 500	500	1	333	240	0.990	1.000	1.000
	500 / 740	1000	2	333	480	1.000	1.000	1.000
	500 / 980	1500	3	333	720	1.000	1.000	1.000
	500 / 1220	2500	4	333	960	1.000	1.000	1.000

K multiplication factor for the number of controls tested, F(*X*+) = frequency of allele *X*,

F(*Y*+) = frequency of allele *Y*.

[a] Total number of cases and controls in the study.

[b] The required number of random controls that need to be tested.

[c] The expected number of cases that have *X*+ in the study.

[d] The expected number of controls that have *X*+ in the study.

To examine the effect of adding controls in a specific stratum in gene-gene interaction studies power calculations were done (table 4). The results indicate that when the risk alleles *X* and *Y* are common (frequency > 0.25), typing additional controls increases the power. This is however not required, as the power without the additional controls is already substantial (>0.869). When one of the risk alleles, *X* and / or *Y*, is more rare (frequency = 0.10), a large increase in power can be obtained by genotyping only a limited number of additional controls. For most cases $K = 2$ is sufficient, unless both risk alleles have a frequency of 0.10. In this case *K* needs to be 3 or 4 and a larger sample of cases and controls needs to be analyzed.

When random controls are added instead of controls carrying one risk allele, the required number of additional tested samples is much higher (table 4). In this case the initial number of controls is multiplied by *K*. The genotyping of only controls which carry the risk allele of the first tested polymorphism (allele *X* in table 4) is consequently more cost efficient.

Discussion

Studying gene-gene interaction is likely the next step in unraveling complex genetic traits. However, very large population-based samples are required that are not always available[23,24]. Here, we show a simple and direct approach to evaluate false-positive findings in gene-gene interaction studies with an example of Alzheimer's disease, PSEN-1 and APOE[5-8]. The first step of the approach is to evaluate cases and controls separately to identify the group(s) in which an interaction is found. Basically, we assume that in gene-gene interaction studies the interaction is explained by association in cases and not in controls. Studying cases only to test for gene-gene interaction has been proposed but often controls are included to allow for estimations of risk[11,12]. If an association between two unlinked genes is found in controls, while cases do not show such a relation, it may be the result of over-stratification of the data.

In this case the second step of the approach, adding controls to the study can be applied.

It should be noted that other reasons for association in controls are not excluded in this manner. Of course, linkage disequilibrium, which may lead to association between the two genes, needs to be excluded *a priori*. Another reason for association may be differential mortality related to the risk alleles. Both mechanisms will have effect in cases and (age-matched) controls, leading to association in both series. No indication for population stratification was found in the Rotterdam study. However, when working in a more mixed population, for example the US population, one may want to consider testing for hidden population stratification first[25-28]. Adding controls may not overcome this problem, nor may it eliminate all variation in the risk estimate for controls, as it was the case in our example of AD.

Extra controls can be added genotyping a random control group, or samples of a specific stratum. In large epidemiological studies where specific genes have been tested already, the addition of controls in a stratum is a more cost-effective alternative. The efficiency of this approach is particularly high when rare risk alleles are studied. The addition of extra controls also has another advantage, namely that the detection power of interactions increases. Therefore, in addition to obtaining a more accurate risk estimate for the control group carrying both risk alleles, the approach also decreases the false-negative rate of the study. Future gene-gene interaction studies intending to find moderate risk factors for complex diseases that require large numbers of genotyped and phenotyped individuals may therefore benefit from this approach.

Acknowledgements

This project was supported by research grants from the Netherlands Organization of Scientific Research (NWO), the Special Research Fund of the

University of Antwerp, the Fund for Scientific Research Flanders (FWO). Bart Dermout is a PhD fellow of the FWO, the Medical Foundation Queen Elisabeth, the Interuniversity Attraction Poles (IUAP) program P5/19 of the Federal Office of Scientific, Technical and Cultural Affairs (OSTC), and the International Alzheimer Research Foundation (IARF), Belgium.

References

1. Baron M. The search for complex disease genes: fault by linkage or fault by association? Mol Psychiatry. 2001;6:143-9.
2. Cardon LR, Bell JI. Association study designs for complex diseases. Nat Rev Genet. 2001;2:91-9.
3. Farrer LA, Cupples LA, Haines JL et al. Effects of age, sex, and ethnicity on the association between apolipoprotein E genotype and Alzheimer's disease. A meta-analysis. APOE and Alzheimer's Disease Meta Analysis Consortium. JAMA. 1997;278:1349-56.
4. Sherrington R, Rogaev EI, Liang Y et al. Cloning of a gene bearing missense mutations in early-onset familial Alzheimer's disease. Nature. 1995;375:754-60.
5. Wragg M, Hutton M, Talbot C. Genetic association between intronic polymorphism in presenilin-1 gene and late-onset Alzheimer's disease. Alzheimer's Disease Collaborative Group. Lancet. 1996;347:509-12.
6. van Duijn CM, Cruts M, Theuns J et al. Genetic association of the presenilin-1 regulatory region with early-onset Alzheimer's disease in a population-based sample. Eur J Hum Genet. 1999;7:801-6.
7. Taddei K, Yang D, Fisher C et al. No association of presenilin-1 intronic polymorphism and Alzheimer's disease in Australia. Neurosci Lett. 1998;246:178-80.
8. Bagli M, Papassotiropoulos A, Schwab SG et al. No association between an intronic polymorphism in the presenilin-1 gene and Alzheimer's disease in a German population. J Neurol Sci. 1999;167:34-6.
9. Goldstein AM, Falk RT, Korczak JF, Lubin JH. Detecting gene-environment interactions using a case-control design. Genet Epidemiol. 1997;14:1085-9.
10. Goldstein AM, Andrieu N. Detection of interaction involving identified genes: available study designs. J Natl Cancer Inst Monogr. 1999;26:49-54.
11. Yang Q, Khoury MJ, Sun F, Flanders WD. Case-only design to measure gene-gene interaction. Epidemiology. 1999;10:167-70.
12. Brennan P. Gene-environment interaction and aetiology of cancer: what does it mean and how can we measure it? Carcinogenesis. 2002;23:381-7.
13. Vandenbroucke JP, Hofman A. Grondslagen der epidemiologie. 6 dr. Maarssen: Elsevier. p. 131. 1999.
14. Miettinen OS. Individual matching with multiple controls in the case of all-or-none response. Biometrics. 1969;22:339-55.
15. Hofman A, Grobbee DE, de Jong PT, van den Ouweland FA. Determinants of disease and disability in the elderly: the Rotterdam Elderly Study. Eur J Epidemiol. 1991;7:403-22.
16. Dermaut B, Roks G, Theuns J et al. Variable expression of presenilin 1 is not a major determinant of risk for late-onset Alzheimer's disease. J Neurol. 2001;248:935-9.

17. Theuns J, Remacle J, Killick R et al. Alzheimer-associated C allele of the promoter polymorphism -22C>T causes a critical neuron-specific decrease of presenilin 1 expression. Hum Mol Genet. 2003;12:869-77.
18. Slooter AJ, Bots ML, Havekes LM et al. apolipoprotein E and carotid artery atherosclerosis: the Rotterdam study. Stroke. 2001;32:1947-52.
19. Das HK, McPherson J, Bruns GA et al. Isolation, characterization, and mapping to chromosome 19 of the human apolipoprotein E gene. J Biol Chem. 1985;260:6240-7.
20. Lambert JC, Mann DM, Harris JM et al. The -48 C/T polymorphism in the presenilin 1 promoter is associated with an increased risk of developing Alzheimer's disease and an increased Abeta load in brain. J Med Genet. 2001;38:353-5.
21. Gordon D, Finch SJ, Nothnagel M, Ott J. Power and sample size calculations for case-control genetic association tests when errors present: application to single nucleotide polymorphisms. Human Heredity 2002;54:22-33
22. Gordon D, Levenstien MA, Finch SJ, Ott J. Errors and linkage disequilibrium interact multiplicatively when computing sample sizes for genetic case-control association studies. Pacific Symposium on Biocomputing: 2003;490-501.
23. Gauderman WJ. Sample size requirements for association studies of gene-gene interaction. Am J Epidemiol. 2002;155:478-84.
24. Yang Q, Khoury MJ, Flanders WD. Sample size requirements in case-only designs to detect gene-environment interaction. Am J Epidemiol. 1997;146:713-20.
25. Pritchard JK, Rosenberg NA. Use of unlinked genetic markers to detect population stratification in association studies. Am J Hum Genet. 1999;65:220-8.
26. Pritchard JK, Stephens M, Rosenberg NA, Donnelly P. Association mapping in structured populations. Am J Hum Genet. 2000;67:170-81.
27. Devlin B, Roeder K. Genomic control for association studies. Biometrics. 1999;55:997-1004.
28. Hoggart CJ, Parra EJ, Shriver MD et al. Control of confounding of genetic associations in stratified populations. Am J Hum Genet. 2003;72:1492-504.