The 7th International Symposium on Operations Research and Its Applications (ISORA'08)
Lijiang, China, October 31–Novemver 3, 2008
Copyright © 2008 ORSC & APORC, pp. 432–439

Multilocus Analysis of Case-control Data to Schizophrenia Based on Measures of Information Discrepancy

Junhua Zhang[1] Weihua Yue[2] Xiang-Sun Zhang[1]

[1] Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing, China
[2] Mental Health Institute, Peking University, Beijing, China

Abstract As one of complex diseases, schizophrenia is considered to possess a complex trait to which genetic, environmental, and epigenetic factors contribute interactively. The genetic analysis of schizophrenia has revealed complex and inconsistent results, making it difficult to draw clear conclusions regarding the impact of specific genes on the disease in diverse human populations. So the need for identification of susceptibility genes especially multilocus interactions for schizophrenia still poses a great and unanswered challenge. In this paper we propose a new approach to characterize the individual and combined contributions of three recently proposed candidate genes (NRG1, G72 and RGS4) in the schizophrenia case-control dataset in the Chinese Han population. The approach is based on **m**easures of **i**nformation **d**iscrepancy (**MID**). First the susceptible loci are selected based on MID by taking into consideration the gene–gene interactions. Then a discrimination algorithm is introduced to assess the effectiveness of the selected multiple loci to predict the disease status, that is, a classification accuracy is obtained by reclassifying the subjects into the case or control groups. Several possible pathogenic locus-locus interactions for schizophrenia are obtained.

Keywords Schizophrenia; Multilocus analysis; Information discrepancy

1 Introduction

Schizophrenia is a severe psychiatric disorder that affects almost 1% of the world's population and accounts for about 2.5% of health-care costs ([1]). It has been reported that schizophrenia has a heritability of about 80% ([2]). Family, twin, and adoption studies suggest that genetic factors play an important role in the etiology of schizophrenia ([3]). However, progress in the search for susceptibility chromosomal loci and schizophrenia-related genes has been slow and unsatisfying, probably because there are multiple genes that may interact with environmental and epigenetic factors to affect susceptibility to the disease. Moreover, several practical problems have impeded the progress of genetic research, such as the lack of a major effect of a gene abnormality on occurrence of schizophrenia and the absence of neuropathological or other biological markers to be used for diagnosis of schizophrenia.

The systematic positional cloning efforts have identified several strong candidate chromosome regions, such as: 1q21-q22, 6p, 8p, 13q, 22q11, etc ([4-9]). Several possible susceptibility genes located in or beside the above chromosome regions, such as the regulator of G-protein signaling-4 (RGS4, 1q21-q22), dystrobrevin-binding protein 1 (DTNBP1, 6p22.3), neuregulin 1 (NRG1, 8p22-p11), G72 (13q34), and catechol-O-methyltransferase (COMT, 22q11.2) were reported to contain single nucleotide polymorphisms (SNPs) and haplotypes associated with schizophrenia ([4-7,10-12]). However, while the evidence implicating the above genes in development of schizophrenia is promising, it is not yet absolutely persuasive considering several failed attempts to show an association. This discrepancy is probably due to differences in test populations and methods ([13-15]). On the other hand, according to the prevailing pathogenic model, schizophrenia is a neurodevelopmental disorder leading to abnormality of synaptic connectivity ([16-18]). Glutamatergic transmission via N-methyl-D-aspartate (NMDA) receptors may be especially involved ([19]). In a recent review, the effects of variation in several genes on schizophrenia, including NRG1, G72, and RGS4 were speculated to influence synaptic plasticity and the cortical microcircuitry ([20]).

A large number of studies ([21-26]) indicate that, variants of NRG1, G72 and RGS4 represent a set of candidate gene polymorphisms associated with schizophrenia susceptibility, and the combined contribution of these genes remains unclear. Recent research comparing the contribution of three candidate genes, G72, NRG1 and DTNBP1, to schizophrenia susceptibility in two demographically distinct familial populations did not support the hypothesis that the three genes interact in influencing susceptibility for the disease ([27]). Here we attempted to characterize the individual and combined contributions of these three recently proposed candidate genes (NRG1, G72 and RGS4) in the schizophrenia case-control dataset in the Chinese Han population. A new approach based on **m**easures of **i**nformation **d**iscrepancy (**MID**) is proposed in this paper. So our approach is called MID approach. MID can be used to identify pathogenic genes for common complex diseases, the main methodology is as followings. First the susceptible loci are selected based on MID by taking into consideration the gene–gene interactions. Then a discrimination algorithm is introduced to assess the effectiveness of the selected multiple loci to predict the disease status, that is, a classification accuracy is obtained by reclassifying the subjects into the case or control groups.

The MID approach is model-free in that it does not assume any particular genetic model and is nonparametric in that it does not estimate any parameters. Moreover, the computation for MID is sparing time and the implementation for it is easy. And through analyzing a real schizophrenia case-control dataset in the Chinese Han population, several possible pathogenic locus-locus interactions for schizophrenia are obtained.

2 Materials and Data

The case-control data used in this paper is from the Mental Health Institute, Peking University. The ethics committee of Peking University Health Science Center approved the study protocol and informed consent for participation was obtained from all subjects. The samples consisted of 147 unrelated schizophrenia cases and 136 unrelated normotensive controls from Chinese Han population. All the schizophrenia cases accord with the standards of the International Classification of Diseases-10 (ICD-10) and the Chinese Classification of Mental Disorders Third Revision (CCMD-3).

Three genes which are recently reported to be candidates to schizophrenia susceptibility are used here. These genes locate on three different chromosomes on which totally thirteen SNP loci were genotyped (**Table 1**).

Table 1. Candidate genes and SNP loci assessed

Gene	SNP locus	Location
NRG1	rs3924999 rs2954041 rs2919390 rs6988339 rs3735774 SNP8NRG221533 SNP8NRG243177	8p22-p11
G72	rs2391191 rs778294 rs947267	13q34
RGS4	Novel-SNP rs12753561 rs10759	1q21-q22

3 Methods

In this section we describe the methods used in this paper. First we select the informative loci based on **m**easures of **i**nformation **d**iscrepancy (MID), so the method is called MID method. Then we introduce an algorithm to reclassify the samples only using the information included in the selected loci (not all the loci). It's worthy to point out that through the latter step we can validate the effectivity of the selected informative loci for schizophrenia to some extent. That is, if the prediction accuracy is high, then it is reasonable to infer that the informative loci are the possible pathogenic loci.

3.1 Selection of The Informative Loci

Let n_1 and n_2 represent the number of the individuals in the case group(called group 1) and the control group(called group 2), respectively. Suppose there are M candidate loci with r_i genotypes at locus i $(1 \leq i \leq M)$.

3.1.1 Selection of Single Informative Loci

For a particular locus with r genotypes, let $(\hat{p}(1,1),\ldots,\hat{p}(1,r))$ and $(\hat{p}(2,1),\ldots,\hat{p}(2,r))$ represent the genotype frequencies in group 1 and group 2, respectively. To measure the worth of a particular locus, an intuitive idea is to measure the amount of information it possesses for discriminating among the groups. So the following MID is used:

$$B = (n_1+n_2)\sum_{k=1}^{2}\sum_{j=1}^{r}\hat{p}(k,j)\ln\frac{\hat{p}(k,j)}{(\hat{p}(1,j)+\hat{p}(2,j))/2}. \tag{1}$$

Originally the MID is introduced by Fang ([28]) to measure the degree of disagreement among multiple information sources. It has been proven that B possesses many good properties, such as non-negativity, symmetry, boundedness, uniform continuity, monotonicity, convexity and so on ([28,29]). Here a particular locus is considered informative if it has high difference B between the case and control groups in genotype frequency over there. And the significance of the difference can be tested by the following hypothesis:

$$H_0: \quad p(1,j) = p(2,j), \quad j = 1,\ldots,r,$$

where $p(k,j)$ is the probability of an individual who belongs to group k and possesses the j-th genotype at the locus $(k = 1,2)$. Under H_0, B is asymptotically distributed as χ^2 with $r-1$ degrees of freedom when $n_1,n_2 \to \infty$ ([30]). Let b denote the observed value of the corresponding MID B, we get the p-value

$$p = P(\chi^2(r-1) \geq b). \tag{2}$$

It is obvious, the smaller p, the more associated the locus is.

3.1.2 Selection of Multiloci Based on Maximum Additional Information

For any particular locus, we examine other loci to investigate which can provide *maximum additional information* to it. For notational convenience, we denote the considered particular locus as locus 1. We look at all locus pairs $(1,Y)$, $2 \leq Y \leq M$. Let $p(k,j,y)$ represent the probability of an individual who belongs to group k and possesses the j-th genotype at locus 1 and the y-th genotype at locus Y. Then the above problem can be investigated through the following hypothesis test

$$H_{1,Y}^{(0)}: \quad p(k,y|j) = p(k|j)p(y|j) \quad (1 \leq j \leq r_1,\ 1 \leq y \leq r_Y,\ k = 1,2),$$

where $p(k,y|j)$ represents the corresponding conditional probability of an individual being of the j-th genotype at locus 1, and similarly for $p(k|j)$ and $p(y|j)$.

To test the null hypothesis $H_{1,Y}^{(0)}$, a statistic $B(Y;1)$, which is similar to the MID in (1), is used:

It is known that under $H_{1,Y}^{(0)}$, $B(Y;1)$ is asymptotically distributed as χ^2 with $r_1(r_Y - 1)$ degrees of freedom when $n_1, n_2 \to \infty$ ([30]). Let $b(Y;1)$ denote the observed value of $B(Y;1)$, now we get the p-value

$$p_{1,Y} = P(\chi^2(r_1(r_Y - 1)) \geq b(Y;1)). \tag{3}$$

And the smaller the $p_{1,Y}$ is, the more additional information the locus Y yields.

Based on the above procedure, the multiloci can be investigated and selected sequentially.

3.2 Discrimination with The Selected Loci

Discrimination is an important problem in statistics science, and it has been greatly applied to many other areas. Here discrimination concerns with classifying any individual into a certain group (case or control). The schizophrenia SNP data is investigated and our goal is to find the possible pathogenic genes. So here we construct an algorithm for discrimination only using the selected loci in aims to validate the cooperation effectivity of them for schizophrenia.

Suppose there are m loci having been selected, and locus i possesses r_i genotypes $(1 \leq i \leq m)$; further, group k has n_k members $(k = 1,2)$. When the cooperation of different loci is considered, we'll obtain a $\gamma(\overset{\triangle}{=} r_1 r_2 \cdots r_m)$-dimensional vector for each group. Let $x_{j_1 j_2 \cdots j_m}^{(k)}$ denote the number of individuals in group k who select the j_i-th genotype at locus i, then $\sum\limits_{j_m=1}^{r_m} \cdots \sum\limits_{j_1=1}^{r_1} x_{j_1 j_2 \cdots j_m}^{(k)} = n_k (k = 1,2)$. We can easily get the genotype frequencies for each group over the combinational permissible genotypes:

$$\hat{p}_{j_1 j_2 \cdots j_m}^{(k)} = \frac{x_{j_1 j_2 \cdots j_m}^{(k)}}{n_k} \quad (1 \leq j_1 \leq r_1, \ldots, 1 \leq j_m \leq r_m;\ k = 1,2). \tag{4}$$

Given any individual who possesses the j_i^*-th genotype at locus i, so his(or her) combinational genotype over the selected m loci is $j_1^* j_2^* \cdots j_m^*$. We compare $\hat{p}_{j_1^* j_2^* \cdots j_m^*}^{(1)}$ with

$\hat{p}^{(2)}_{j_1^* j_2^* \cdots j_m^*}$, and classify the individual into the group which has the larger corresponding frequency. It is easy to imagine that the larger the frequency, the greater the possibility that the individual belongs to the corresponding group.

4 Results

Here the schizophrenia case-control data described in the section MATERIALS AND DATA, for which each SNP locus is of a bi-allele(i.e., tri-genotype), are analyzed by the proposed MID method.

When each of the three genes are investigated separately, two significant informative loci are selected (**Table 2**). It is worthy to notice that the significance level for every gene has been with Bonferroni correction considering the multiple testing. For example, for the gene NRG1 with seven SNP loci, the significance level is $0.05/7 \approx 0.0071$. In Table 2, for each significant informative locus the corresponding p-value computed by (2) is also given.

Table 2. Significant informative single loci and the corresponding p-values

Gene	SNP locus	*p*-value
NRG1	rs3924999	0.0056
RGS4	rs10759	0.0057

When gene-gene interaction comes into consideration for schizophrenia susceptibility, some loci combinations are obtained (**Table 3**).

Table 3. Multiloci combinations with significant additional information

loci combination	*p*-value
rs3924999 / *rs6988339*	*9.9122e-004*
rs3924999 / *rs2391191*	*0.0011*
rs3924999 / *rs10759*	*0.0027*
rs2954041 / *rs3924999*	*0.0034*
rs2919390 / *rs3924999*	*0.0014*
rs6988339 / *rs3924999*	*7.5542e-005*
SNP8NRG221533 / *rs3924999*	*9.3020e-004*
rs2391191 / *rs3924999*	*2.3117e-004*
rs778294 / *rs3924999*	*6.4806e-005*
rs778294 / *rs10759*	*0.0032*
rs12753561 / *rs10759*	*0.0020*
rs10759 / *rs3924999*	*0.0028*
SNP8NRG221533 / rs3924999 / *rs10759*	*0.0090*

For each such combination, the latter locus provides significant additional information for the previous loci. Now using the Bonferroni correction for multiple testing, the significance level for the second additional locus is $0.05/12 \approx 0.0042$. In order to get any possible 3-locus combination, here the significance level for the third additional locus is set as $0.10/11 \approx 0.0091$.

To evaluate the effectiveness of the selected loci or loci combinations to distinguish the two different groups of schizophrenia patients and the healthy people, we use the discrimination method introduced above to reclassify the subjects from the complete dataset.

The classification accuracies by using the significant informative single loci as well as the loci combinations with significant additional information are listed in **Table 4**.

Table 4. The classification accuracies by using some different SNP loci or loci combinations

locus or loci combination	classification accuracy
rs3924999	57.95%
rs10759	55.48%
rs3924999/rs6988339	61.48%
rs3924999/rs2391191	62.54%
rs3924999/rs10759	**63.25%**
rs2954041/rs3924999	58.66%
rs2919390/rs3924999	**64.31%**
SNP8NRG221533/rs3924999	62.19%
rs778294/rs3924999	60.07%
rs778294/rs10759	57.95%
rs12753561/rs10759	62.19%
SNP8NRG221533/rs3924999/rs10759	**65.72%**

From Table 4 we can see that the classification accuracy is a little lower when the significant informative single locus NRG1*rs3924999 or RGS4*rs10759 is used. However the accuracies are increased when the loci combinations are used. This indicates that schizophrenia is more possible caused by locus-locus or gene-gene interactions rather than a single locus or a single gene. Speaking in details, the several higher accuracies corresponding to NRG1*SNP8NRG221533 /NRG1*rs3924999/RGS4*rs10759 (65.72%), NRG1*rs2919390/NRG1*rs3924999 (64.31%) and NRG1*rs3924999/RGS4*rs10759 (63.25%), respectively. So we confer that the locus-locus interactions between genes NRG1 and RGS4 as well as the interaction within gene NRG1 may operate mainly for schizophrenia susceptibility for Chinese Han population.

5 Conclusion

In this paper a new method is proposed to characterize the individual and combined contributions of multiple genes to complex diseases using case-control dataset. The method is based on measures of information discrepancy, so it is called MID method. Here we use it to analyze the dataset from Chinese Han population on the three recently proposed candidate genes (NRG1, G72 and RGS4) to schizophrenia, several possible pathogenic loci combinations are identified. It is worthy to point out that the MID method is model-free in that it does not assume any particular genetic model and does not estimate any parameters, that is, the implementation for it is quite easy.

Although many approaches have been developed before in the research of complex diseases, we hope that our MID method will be a helpful complementarity to this field. And we expect that this new method will be employed with promising results in the exploration of many puzzling complex diseases.

References

[1] Meltzer D. Perspective and the measurement of costs and benefits for cost-effectiveness analysis in schizophrenia. J Clin Psychiatry. 1999; Suppl 3: 32-35.

[2] Owen M.J., O'Donovan M, Gottesman II (ed). (2003) Psychiatric genetics and genomics. Oxford, pp 247-266.

[3] Kendler K.S. and Diehl S.R. (1993) The genetics of schizophrenia: A current, genetic-epidemiologic perspective. *Schizophr Bull*, **19**, pp 261-285.

[4] Brzustowicz LM, Hodgkinson KA, Chow EW, et al. Location of a major susceptibility locus for familial schizophrenia on chromosome 1q21-22. Science. 2000; 288: 678-682.

[5] Straub RE, Jiang Y, MacLean CJ, et al. Genetic variation in the 6p22.3 gene DTNBP1, the human ortholog of the mouse dysbindin gene, is associated with schizophrenia. Am J Hum Genet. 2002; 71: 337-348.

[6] Stefansson H, Sigurdsson E, Steinthorsdottir V, et al. Neuregulin 1 and susceptibility to schizophrenia. Am J Hum Genet. 2002; 71: 877-892.

[7] Chumakov I, Blumenfeld M, Guerassimenko O, et al. Genetic and physiological data implicating the new human gene G72 and the gene for D-amino acid oxidase in schizophrenia. Proc Natl Acad Sci USA. 2002; 99: 13675-13680.

[8] Liu H, Heath SC, Sobin C, et al. Genetic variation at the22q11 PRODH2/DGCR6 locus presents an unusual pattern and increases susceptibility to schizophrenia. Proc Natl Acad Sci USA. 2002; 99: 3717-3722.

[9] Liu H, Abecasis GR, Heath SC, et al. Genetic variation in the 22q11 locus and susceptibility to schizophrenia. Proc Natl Acad Sci USA. 2002; 99: 16859-16864.

[10] Jacquet H, Raux G, Thibaut F, et al. PRODH mutations and hyperprolinemia in a subset of schizophrenic patients. Hum Mol Genet. 2002; 11: 2243-2249.

[11] Stefansson H, Sarginson J, Kong A, et al. Association of neuregulin 1 with schizophrenia confirmed in a Scottish population. Am J Hum Genet. 2002; 72: 83-87.

[12] Shifman S, Bronstein M, Sternfeld M, et al. A highly significant association between a COMT haplotype and schizophrenia. Am J Hum Genet. 2002; 71: 1296-1302.

[13] Iwata N, Suzuki T, Ikeda M, et al. No association with the neuregulin 1 haplotype to Japanese schizophrenia. Mol Psychiatry. 2003; 9: 126-127.

[14] Thiselton DL, Webb BT, Neale BM, et al. No evidence for linkage or association of neuregulin-1 (NRG1) with disease in the Irish study of high-density schizophrenia families (ISHDSF). Mol Psychiatry. 2004; 9: 777-783.

[15] Mulle JG, Chowdari KV, Nimgaonkar V, Chakravarti A. No evidence for association to the G72/G30 locus in an independent sample of schizophrenia families. Mol Psychiatry. 2005; (in press)

[16] Weinberger DR. Implications of normal brain development for the pathogenesis of schizophrenia. Arch Gen Psychiatry. 1987; 44: 660-669.

[17] Lewis DA, Levitt P. Schizophrenia as a disorder of neurodevelopment. Annu Rev Neurosci. 2002; 25: 409-432.

[18] Harrison PJ. The neuropathology of schizophrenia: a critical review of the data and their interpretation. Brain. 1999; 122: 593-624.

[19] Tsai G, Coyle JT. Glutamatergic mechanisms in schizophrenia. Annu Rev Phamacol Toxicol. 2002; 42: 165-179.

[20] Harrison PJ, Weinberger DR. Schizophrenia genes, gene expression, and neuropathology: on the matter of their convergence. Mol Psychiatry. 2005; 10: 40-68.

[21] Buonanno A, Fischbach GD. Neuregulin and ErbB receptor signaling pathways in the nervous system. Curr Opin Neurobiol. 2001; 11: 287-296.

[22] Mothet JP, Parent AT, Wolosker H, et al. D-serine is an endogenous ligand for the lycine site of the N-methyl-D-aspartate receptor. Proc Natl Acad Sci USA. 2000; 97: 4926-4931.

[23] Korostishevsky M, Kaganovich M, Cholostoy A, et al. Is the G72/G30 locus associated with schizophrenia? single nucleotide polymorphisms, haplotypes, and gene expression analysis. Biol Psychiatry. 2004; 56: 169-176.

[24] De Blasi A, Conn PJ, Pin J, Nicoletti F. Molecular determinants of metabotropic glutamate receptor signaling. Trends Pharmacol Sci. 2001; 22: 114 -120.

[25] Thaminy S, Auerbach D, Arnoldo A, Stagljar I. Identification of novel ErbB3 interacting factors using the split-ubiquitin membrane yeast two-hybrid system. Genome Res. 2003; 13: 1744 -1753.

[26] Corfas G, Roy K, Buxbaum JD. Neuregulin 1-erbB signaling and the molecular/cellular basis of schizophrenia. Nat Neurosci. 2004; 7: 575 -580.

[27] Hall D, Gogos JA, Karayiorgou M. The contribution of three strong candidate schizophrenia susceptibility genes in demographically distinct populations. Genes Brian Behav. 2004; 3: 240-248.

[28] Fang, W.W. The disagreement degree of multi–person judgments in additive structure. Mathematical social sciences. 1994; 28(2): 85–111.

[29] Fang, W.W. The characterization of a measure of information discrepancy. Information Science. 2000; 125: 207–232.

[30] Zhang J.H. and Fang W.W. A new approach of information discrepancy to analysis of questionnaire data. Communications in Statistics — Theory and Methods. 2003; 32(2): 435–457.