Statistical Science
2011, Vol. 26, No. 1, 116–129
DOI: 10.1214/11-STS353

Statistical Analysis in Genetic Studies of Mental Illnesses

Heping Zhang

Abstract. Identifying the risk factors for mental illnesses is of significant public health importance. Diagnosis, stigma associated with mental illnesses, comorbidity, and complex etiologies, among others, make it very challenging to study mental disorders. Genetic studies of mental illnesses date back at least a century ago, beginning with descriptive studies based on Mendelian laws of inheritance. A variety of study designs including twin studies, family studies, linkage analysis, and more recently, genomewide association studies have been employed to study the genetics of mental illnesses, or complex diseases in general. In this paper, I will present the challenges and methods from a statistical perspective and focus on genetic association studies.

Key words and phrases: Comorbidity, covariate adjusted association test, FBAT, Kendall's tau, multiple traits, ordinal traits.

1. INTRODUCTION

Mental illnesses affect the health and well-being of all populations and all ages. Schizophrenia—a chronic, severe, and disabling brain disorder—is one of these mental illnesses, affecting about 1.1 percent of the U.S. population age 18 and older in a given year. People with schizophrenia sometimes hear voices others do not hear, believe that others are broadcasting their thoughts to the world, or become convinced that others are plotting to harm them. These experiences can make them fearful and withdrawn and cause difficulties when these people try to have relationships with others (http://www.nimh.nih.gov). Emil Kraepelin (1856–1926) described "Dementia Praecox" as an inherited disorder in his influential "Textbook of Psychiatry" (1899). Dementia Praecox, coined "schizophrenia," was first used by Arnold Pick (1851–1924)—a professor of psychiatry at the German branch of Charles University in Prague—to describe a patient with a psychotic disorder resembling hebephrenia in 1891.

Nearly a century ago, Cannon and Rosanoff (1911) made an attempt to understand whether there are any forms of nervous and mental diseases that are transmitted from generation to generation in concordance with Mendelian laws. They examined the families of 11 neuropathetic patients, which are now referred to as probands in pedigrees. Using Mendelian laws as their theoretical expectation, they concluded that the neuropathetic make-up is recessive to normal. Although the report was indeed "preliminary," a few things are noteworthy. First, they noted that "any form of insanity or even all the forms of hereditary insanity do not constitute an independent hereditary character." This raised an early sign of the complexity associated with studying mental disorders compared to the characterization of the disorders and their comorbidity. Here, comorbidity refers to more than one disease condition in the same patient. Second, they remarked "should larger accumulations of such data in the future give similar results, we shall be able" to confirm their result. The requirement for more samples and replication is another challenge in studies of complex diseases. Last, but not the least, while they said "let us test, ..., the hypothesis ..." they did not mean a statistical test. However, the idea of the χ^2-test is evident.

Despite this early work, it was not until the 1960s that the researchers began to use scientifically rigorous designs and methods to study the inheritance of mental illnesses. For example, the key idea in adoption studies lies in the belief that any links between an adopted child and the biological parents are attributable to genetics, and any links between that child and adoptive parents can be attributed to environment (Plomin

Heping Zhang is Professor, Yale School of Public Health, Yale University, 60 College Street, New Heaven, Connecticut 06520-8034, USA (e-mail: heping.zhang@yale.edu).

et al., 1997). This enables us to separate the confounding environment (i.e., a family) from genetic contribution. Consequently, there are two strategies in adoption studies. One approach compares the risk of developing schizophrenia in the adopted children of schizophrenic parents to the risk of adopted children whose parents do not have schizophrenia. Several studies including Heston (1966), Rosenthal (1972) and Tienari (1991) used this approach to study schizophrenia. Each study found an elevated risk in adopted-away children of schizophrenic parents, supporting the role of genetics in the transmission of schizophrenia. The origin of this approach is the schizophrenic parents. Another approach backtracks from adopted children who have developed schizophrenia and compares the risks of schizophrenia in their adoptive and biological families. Kety, Rosenthal and Wender (1978) and others found that the risk was significantly higher in the biological relatives than in the adoptive families, again underscoring the role of genetics as a risk factor.

While these schizophrenia adoption studies are influential in understanding the role of genetics in mental disorders, the majority of the genetic factors associated with mental disorders are based on family and twin studies. By comparing the concordance in the risk between identical (monozygotic) and fraternal (dizygotic) twins, twin studies arguably provide the most compelling results about genetic and environmental effects. For example, the concordance in monozygotic twins for Tourette's syndrome, a complex disorder characterized by repetitive, sudden and involuntary movements or noises called tics, was reported to be about 50% whereas it is less than 10% in dizygotic twins.

Twin studies are most helpful in demonstrating the magnitude of genetic effect, but they do not provide insight into the inheritance pattern of a condition. Thus, family studies can offer information that twin studies cannot. Thus, Cannon and Rosanoff (1911) employed a small-scale, simple family study. Using the Mendelian laws, not only might we find evidence of genetics, but also infer the mode of transmission, as Cannon and Rosanoff (1911) concluded for the heredity of insanity.

Although twin and family studies continue to be useful for understanding the genetics of complex diseases, different studies are needed to locate a specific gene on a chromosome that may underlie the disease. Gene mapping in humans through linkage analysis emerged in the 1930s, but it was Morton (1955) who laid the foundation for the methodology. It was only during the 1970s and 1980s, when the Elston–Steward (1971) algorithm was developed and implemented (Ott, 1974), that the method thrived as a common tool of genetic studies. These initial and subsequent developments allowed for linkage analyses of multiple markers simultaneously. In light of the sheer number of genes and that we do not know which specific gene we are looking for, we typically genotype 300 to 400 "landmarks" that cover the 22 pairs of autosomes and the X chromosome. By inferring the transmission patterns of these markers, then linking them to the disease status, we can obtain information about the most probable region where the gene of interest resides.

While linkage studies have had some successes (e.g., BRCA1), they have generated many more premature excitements. In the late 1980s, two particular studies attracted significant public attention after they reported that bipolar affective disorders were linked to DNA markers on chromosome 11, and that a susceptibility locus for schizophrenia was located on chromosome 5. Unfortunately, these findings were not replicated. Replications in genetic studies of mental disorders do not come easily. For example, Abelson et al. (2005) identified mutations involving the SLITRK1 gene (13q31.1) in a small number of people with Tourette's syndrome. However, most people with Tourette's syndrome do not have a mutation in the SLITRK1 gene. Because the mutations were reported in so few people with this condition, the association of the SLITRK1 gene with this disorder could not be confirmed. In fact, Scharf et al. (2008) reported a lack of the association between SLITRK1var321 and Tourette's syndrome in a large family-based sample.

Various reasons have been suggested to explain the difficulties detouring progress in genetic studies using linkage analysis. A key concept underlying linkage analysis is the recombination fraction, which reflects the distance between any two markers, such as a DNA marker and the disease locus. There may be limited information in the data, however, diminishing the power of the linkage study. Furthermore, complex diseases are polygenic, involving multiple genes (Carter and Chung, 1980). Linkage analyses, however, are generally under the assumption of one major gene. Additionally, heterogeneity in the diagnosis and comorbidity of mental illnesses make linkage analysis considerably more difficult, if even possible at all.

Many investigators have adopted association analyses to take advantage of the advent of high-throughput genotyping technologies. Recent efforts have identified genes that contribute to a number of complex hu-

man traits using the ultra-dense genetic markers (Arking et al., 2006; Klein et al., 2005; Duerr et al., 2006; Chen et al., 2007). Trios (one affected offspring and two parents) have been an effective design for association studies, particularly with the development of the elegant transmission/disequilibrium test (TDT) (Spielman, McGinnis and Ewens, 1993). The central idea of this test is that each affected child serves as his or her own matched case and control. This acts to control for all potential confounding issues and examines alleles that both are and are not transmitted from the parents. In the absence of association between the affective status and the gene, the distributions of the transmitted and non-transmitted alleles are expected to be the same. Deviations in distribution as evaluated by a χ^2-test indicate the existence of association. Trios are the simplest example of nuclear family, but when other siblings are available, the trio design is not cost-effective. As a result, family-based association tests (FBAT) including sibships (Spielman and Ewens, 1998; Horvath and Laird, 1998; Knapp, 1999), nuclear families (Weinberg, 1999; Lunetta et al., 2000; Rabinowitz and Laird, 2000) and general pedigrees (Martin, Monks, Warren and Kaplan, 2000) have been developed.

Another restriction in the use of trios is the requirement of defining the affective status of a disease. Consequently, association tests have been proposed for quantitative traits (Allison, 1997; Rabinowitz, 1997), traits with distribution belonging to an exponential family (Liu, Tritchler and Bull, 2002), ordinal traits (Zhang, Wang and Ye, 2006; Wang, Ye and Zhang, 2006) and multiple traits (Lange et al., 2003; Zhang, Liu and Wang, 2010).

Since the early success in identifying the complement factor H polymorphism in age-related macular degeneration (Klein et al., 2005), case-control association studies have intensified, and many genetic variants have been identified and catalogued (Hindorff et al., 2009). Despite the enormous investment, the intense attention to the genetics of diseases, the rapid improvement in technology, and the increasingly large sample sizes in many studies, it remains challenging to identify disease genes, especially those underlying mental illnesses. Some of the common genetic variants that have been identified for complex diseases only account for a small portion of the genetic risk, which may vary across populations (Goldstein, 2009). For example, Kopp et al. (2008) and Kao et al. (2008) identified several variations in the MYH9 gene as major contributors to excess risk of kidney disease among African-Americans. They found that 60 percent of African-Americans carry the risk variants as opposed to 4 percent of white Americans.

Technology will continue to improve and the amount of genetic data will increase. The purpose of this article is to review some of the progress from a statistical perspective and discuss some of the potential challenges. Obviously, it would take volumes or series to do justice to all of the work in statistical genetics. Instead of taking on that impossible task, this article is oriented toward the publications directly related to my own recent work.

2. METHODS

Since 1952, the American Psychiatric Association has published four editions of the Diagnostic and Statistical Manual of Mental Disorders (DSM) and plans to release its fifth edition in 2013. While widely used, the use and development of the DSM has not gone without controversy and criticism. Unlike diseases for which the diagnoses are well accepted by physicians and patients, such as cancer, the diagnosis of mental disorders must reflect biological factors (e.g., gender and racial disparities), non-biological factors such as culture that are not specific to one person, and it also must reflect the natural variation within the same person.

2.1 Ordinal Traits

It is clear from the above discussion that a simple dichotomous diagnosis (e.g., yes or no), or a well-distributed continuous trait, is unlikely to characterize the state of mental disorders. In fact, the questions used in the diagnosis of mental disorders, such as DSM-IV, are usually posed in terms of severity or frequency, and hence in an ordinal scale.

Statistical methods for genetic analysis are well established for both quantitative (continuous) and binary traits (see, e.g., Blackwelder and Elston, 1985; Goldgar, 1990; Schork, 1993; Amos, 1994; Risch and Zhang, 1995; Kruglyak et al., 1996; Blangero and Almasy, 1997; Ott, 1999). While there has been some progress in the analysis of ordinal traits (e.g., Heath et al., 2002; Steinke, Borish and Rosenwasser, 2003; Vergne et al., 2003; Zhang, Feng and Zhu, 2003; Feng, Leckman and Zhang, 2004; Zhang, Liu and Wang, 2010), especially in plant science (Rao and Xu, 1998; Xu and Xu, 2006), insufficient attention has been paid to addressing the unique challenges of analyzing ordinal traits. Some researchers have recognized that it

is difficult to conduct genetic analyses of ordinal traits because such traits cannot be directly characterized by a linear function of genetic and environmental effects (Rao and Xu, 1998). To fill in this methodological gap, we have made a systematic effort to develop statistical methods for segregation analysis (Zhang, Feng and Zhu, 2003), linkage analysis (Feng, Leckman and Zhang, 2004) and association analysis (Zhang, Wang and Ye, 2006) of ordinal traits (for family studies and case-control studies).

2.1.1 *Analysis of family data*. Long before the era of genomics, researchers collected data in families, also called pedigrees as illustrated in Figure 1. Although the ascertainment process for families varies, Figure 1 depicts a representative three-generation pedigree. The proband is the first person who enters into the study according to defined inclusion and exclusion criteria: such criteria are related to the disease of interest. Other members of the proband's family are included and directly or indirectly assessed, depending on the circumstance. The key idea in analyzing family data is that if a gene is a major driving force behind a disease, a trace in the concordance of diseases in family members would reflect the transmission pattern of a gene under the Mendelian laws. This is the fundamental concept that Cannon and Rosanoff (1911) employed. This type of analysis is referred to as segregation analysis.

The Elston–Stewart (1971) algorithm set up the quintessential framework to analyze data from general pedigrees through a technique called peeling. The main complication in analyzing pedigree data is the complex relationship among family members, making it difficult to express the likelihood function in an easily computed form. The peeling algorithm makes use of the conditional independence embedded in the pedigree resulting from the Mendelian laws, and so peels off the complete likelihood function into smaller pieces before putting them back together.

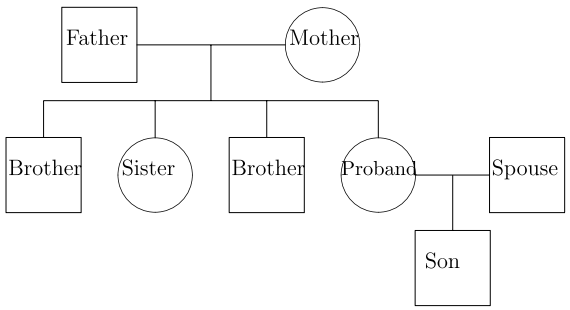

FIG. 1. *A three-generation pedigree.*

Other methods have been relatively recently developed using the concept of latent random variables (Hopper, 1989; Babiker and Cuzick, 1994; Li and Thompson, 1997; Siegmund and McKnight, 1998; Zhang and Merikangas, 2000), which are closely related to the classic ousiotype models of Cannings, Thompson and Skolnick (1978) in pedigree analysis. The basic idea is to use latent variables to represent the contribution of unobserved factors including a major gene, residual genetic factors and common environmental factors. As discussed by Zhang and Merikangas (2000), the computation involving pedigrees is similar to the peeling algorithm. Advantages of using latent variable based models are that the interactions between underlying genetic effects and the observed covariates (e.g., demographic variables) can be considered. Additionally, more relevant to this article, we can accommodate ordinal traits in the latent variable framework.

2.1.2 *A latent variable model*. We follow the notation of Zhang and Merikangas (2000) and Zhang, Feng and Zhu (2003). Consider a trait, Y, that takes an ordinal value of $0, 1, \ldots, K$. Let $\mathbf{x}$ be a p-vector of covariates that is also available for each study subject. Three types of latent random variables U_1^i, U_2^i and U_3^i are introduced within family i to represent, respectively, (a) common, unmeasured environmental factors; (b) genetic susceptibility of the family founders (a founder refers to a subject whose parents are not a part of the observed pedigree, e.g., father, mother and spouse in Figure 1); and (c) the transmission of susceptibility genes from a parent to an offspring.

The concept of latent variables is straightforward, but the interesting and difficult part lies in the specification of their distributions. They need to be interpretable and convenient. The following are the assumptions that we found useful:

- U_1^i follows Bernoulli distributions $P\{U_1^i = 1\} = 1 - P\{U_1^i = 0\} = \theta_1$, where θ_1 is an unknown parameter.
- $U_2^i = (U_{2,1}^i, U_{2,2}^i, \ldots, U_{2,2n_i-1}^i, U_{2,2n_i}^i)'$, where n_i is the size of pedigree i. Here, $P\{U_{2,2j-1}^i = 1\} = 1 - P\{U_{2,2j}^i = 0\} = \theta_2$ when $U_{2,2j-1}^i$ and $U_{2,2j}^i$ are the U_2^i-variables of a founder.
- $U_3^i = (U_{3,1}^i, \ldots, U_{3,s_i}^i)'$. According to the Mendelian laws, $P\{U_{3,j}^i = 1\} = P\{U_{3,j}^i = 0\} = \frac{1}{2}$, $j = 1, \ldots, s_i$, and s_i is the number of parent-offspring pairs in family i. U_3^i facilitates the transmission of U_2^i-variables from the founders to the offspring. For example, if a parent of subject j has U_2^i-variables, $U_{2,2k-1}^i$ and $U_{2,2k}^i$, and the U_3^i-variable for this

parent-offspring pair is $U^i_{3,l}$, then one of subject j's U^i_2-variables is $U^i_{2,2j-1} = U^i_{2,2k-1}U^i_{3,l} + U^i_{2,2k}(1 - U^i_{3,l})$.

- All latent variables are independent.

U^i_1 is a simple "switch" indicating the presence or absence of a shared environment factor within family i. U^i_2 is assigned independently to each of founders who are the source for any gene to enter into a family, and thus mimics the transmission of a single major susceptibility locus with alleles A and a of frequencies θ_2 and $1 - \theta_2$, respectively.

Conditional on all of the latent variables, denoted by U^i, within family i, the probability distribution for member j is assumed to be

$$P\{Y^i_j \le k|U^i\} = \frac{\exp(\mathbf{x}^i_j\beta + \alpha_k + \mathbf{a}^i_j\gamma)}{1 + \exp(\mathbf{x}^i_j\beta + \alpha_k + \mathbf{a}^i_j\gamma)}, \quad k = 0, \ldots, K-1, \tag{2.1}$$

where $\mathbf{a}^i_j = (U^i_1, U^i_{2,2j-1} + U^i_{2,2j}, U^i_{2,2j-1}U^i_{2,2j})^T$, and β and γ are p- and 3-vectors of parameters. The α_k is the trait level dependent intercept, $k = 0, \ldots, K-1$.

As Zhang, Feng and Zhu (2003) pointed out, the β parameters measure the strength of association between the trait and the covariates, conditional on the latent variables. The γ parameters indicate the familial and genetic contributions to the trait. The mode of inheritance can be inferred from γ. For example, $\gamma_2 = 0$ and $\gamma_3 \neq 0$ suggests a recessive effect.

The likelihood function can be derived from (2.1). Due to the presence of latent variables, the EM algorithm (Dempster, Laird and Rubin, 1977) is the most convenient choice for parameter estimation (Guo and Thompson, 1992; Zhang and Merikangas, 2000; Zhang, Feng and Zhu, 2003). Although Zhang and Merikangas (2000) and Zhang, Feng and Zhu (2003) presented an effective solution (e.g., a modified likelihood), we should note that the lack of concavity in the likelihood function makes it a challenging task to find the maximum likelihood estimates of the model parameters. In addition, the θ's and γ's are not fully identifiable. The identifiability issue not only causes computational problems, but also presents theoretical challenges in statistical inference. Another important, yet understudied, issue is the validation of the assumptions on the distributions of the latent variables.

But, how useful is the latent variable model (2.1)? First, it provides a regression framework to assess familial aggregation and genetic contribution, and possibly interactions between measured covariates and latent factors. Using data from a family study of substance use (Merikangas et al., 1998), Zhang and Merikangas (2000) were able to present extremely significant evidence of familial aggregation p-value $< 10^{-9}$ for alcohol dependence. This study additionally demonstrated that transmission does not follow a major locus pattern. In retrospect, their findings predicted the difficulty of identifying major genes associated with alcoholism. In addition, Zhang and Merikangas (2000) presented simulation examples to delineate when the absence of latent variables in (2.1) affects the estimates of the effects by the measured covariates. For example, hypothetically, if the greater presence of females in a family has an impact on the well-being of the family, ignoring the familial latent variables is likely to result in a biased estimate of the sex difference.

Not only is it important to include the latent factors, but also it is important to adjust for covariates. To further illustrate this point, Zhang, Feng and Zhu (2003) reported the following simulation. Ten thousand data sets were generated from model (2.1) with $\theta_1 = 0.3$, β chosen from 0, 1, 5 or 10, γ_1 from 0, 1 or 2, $\alpha_0 = -1$ and $\alpha_1 = 1$. To focus on the difference of having or not having covariates, they set $\gamma_2 = \gamma_3 = 0$. Each data set consists of 200 families with 7 family members (similar to Figure 1). One covariate x was generated as follows. For family i, U^i_1 were generated according to whether a random number r_{i1} from the uniform(0, 1) was greater than 0.3 or not. For member j in family i, an independent random number r_{ij2} from the uniform(0, 1) was generated. Then, $x_{ij} = 0.9r_{ij2} + 0.2r_{i1}$.

To evaluate the performance of the test statistic, the covariate was deliberately ignored in the test. When $\beta = 0$, the covariate played no role in the data generating process. The row corresponding to $\beta = 0$ in Table 1 displays the p-value (the column corresponding to $\gamma_1 = 0$) and the power for two values of γ_1 (1 or 2).

TABLE 1

The probability estimates of rejecting $\gamma_1 = 0$ at the significance level of 0.05. *The covariate is omitted from the testing despite the fact that its coefficient β may not be zero*

	$\gamma_1 = 0$	$\gamma_1 = 1$	$\gamma_1 = 2$
$\beta = 0$	0.0494	0.9503	1.0
$\beta = 1$	0.0534	0.9843	1.0
$\beta = 5$	0.1667	0.9971	1.0
$\beta = 10$	0.3828	0.9890	1.0

When $\beta \neq 0$, the covariate plays a role in the data generating model. The data in Table 1 reveal the consequence of ignoring the covariate, which is more severe when the effect of the covariate is greater.

2.1.3 *Linkage analysis*. While linkage analysis has a long history, it only became a common practice after the availability of several convenient computing programs (Ott, 1974; Kruglyak et al., 1996; Almasy and Blangero, 1998). For statisticians, some of the common terminologies in linkage analysis are puzzling, including the so-called LOD-score method and nonparametric method.

Morton (1955) first introduced the term "LOD-score." LOD stands for "the logarithm (base 10) of odds." The "odds" is a probability ratio, or likelihood ratio, of the probability under an alternative hypothesis to the probability under the null hypothesis. The LOD-score method is essentially a log-likelihood ratio test with two fundamental differences: (a) the use of the base 10 logarithm versus the natural logarithm; (b) the log-likelihood ratio statistic has a multiplier of 2 conforming to a χ^2 distribution under certain regularity conditions.

Specifically, the LOD-score is the log(base 10)-ratio of the likelihood when the recombination fraction is less than $1/2$ (i.e., two loci are not on the same chromosome, or called unlinked), to the likelihood when the recombination fraction is $1/2$ (no linkage). The recombination fraction is the frequency that a chromosomal crossover occurs between two loci (or genes) during meiosis; 1% of combination frequency is termed the distance of one centimorgan (cM) in a genetic linkage map. Because the LOD-score is in base 10, a score of 3 indicates 1000 to 1 odds in favor of the linkage, which is the conventional threshold for declaring the evidence for linkage. If we convert a LOD-score of 3 into the standard log-likelihood ratio statistic, it yields a p-value of 2×10^{-4} under χ^2_1. By Bonnferoni correction, it corresponds to a genomewide p-value of 0.05 for 250 markers. This number is in the range for the number of microsatellites used in typical linkage studies.

In order to compute the LOD-score, we first need a number of parameters that determine the likelihood for a given recombination fraction. Then use the maximum likelihood over the recombination fraction for the likelihood under the alternative hypothesis. The parameters that are required include the mode of inheritance, penetrance, and disease allele frequency. These parameters are generally unknown and difficult to estimate for complex diseases including mental illnesses. For example, using segregation analysis (see Section 2.1.1) Pauls and Leckman (1986) examined specific genetic hypotheses about the mode of transmission of Gilles de la Tourette's syndrome, by performing segregation analyses in 30 nuclear families (two-generation pedigrees). They concluded that Tourette's syndrome is inherited as an autosomal dominant trait (one copy of the abnormal allele is sufficient to cause the disease). The penetrance (the probability of having the disease for a given genotype) was reported at 0.71 in males and 1.0 in females with at least one abnormal allele. After several decades of research, no major genetic variant has been identified for Tourette's syndrome, and most likely this syndrome involves multiple genes, interacting with environmental factors. This reality makes it difficult to infer the mode of inheritance, penetrance, and disease allele frequency, and conceptually, this may not make sense for complex diseases (non-Mendelian inheritance).

This difficulty is somewhat alleviated since the LOD-score method has been found to work reasonably well (e.g., Abreu, Greenberg and Hodge, 1999) under various parameter settings. There have been some efforts to improve the robustness of the method (Gastwirth, 1966, 1985; Whittemore, 1996). See Zheng et al. (2009) for a thorough review. Existing methods do not extend to the case of ordinal traits. The effectiveness of the robust methods remains to be studied. Naturally, nonparametric linkage methods have been developed to avoid specification of the genetic model parameters. In statistics, "nonparametric" methods typically refers to distribution-free methods such as rank-based tests and methods based on the empirical distribution. In linkage analysis, however, "nonparametric" does not mean "distribution-free," but instead refers to the replacement of true genetic model parameters with the parameters of inheritance of markers, hypothesized to be close to the disease locus. Thus, with nonparametric linkage methods, we still need to compute the likelihood. Two core algorithms are used to compute the likelihood: the Elston–Steward algorithm (1971) and the Lander–Green (1987) algorithm. As previously discussed, the Elston–Steward algorithm (1971) is a peeling algorithm that makes the computation in a large pedigree feasible by splitting it into small pieces. This algorithm was implemented in early versions of linkage analysis programs (e.g., LIPED and LINKAGE); computational time increased linearly in family size, but exponentially with the number of loci. More recent programs (e.g., GENEHUNTER) use the Lander–Green (1987) algorithm that has first-order complexity

in the number of loci, but unfortunately exponential in the family size. Although Markov chain Monte Carlo methods have been used to accommodate linkage analysis of large families and a large number of markers (Guo and Thompson, 1992), in practice, one may have to break large pedigrees apart in order to run programs such as GENEHUNTER.

We should note that there had not been a linkage analysis program to handle ordinal traits until the release of LOT (Zhang et al., 2008). Typically, the methods for linkage analysis can be divided into two main steps; only the second step involves the trait (Kruglyak et al., 1996). The first step infers how genetic information travels in a family as represented by the so-called "inheritance vector."

We will use the pedigree in Figure 1 to illustrate this concept. The two parents and spouse are the founders of the family, meaning that their parents are not in the current pedigree. The four siblings and the child are nonfounders. The inheritance pattern at marker locus t is completely described by an inheritance vector $v(t) = (v_1, v_2, v_3, v_4, \ldots, v_9, v_{10})'$. In other words, we devote two elements for every nonfounder. The founders are not included because they are the sources of the genes in the family and the inheritance vector is conditional on their genes. The paired elements describe the outcomes of the paternal and maternal meioses transmitted to the nonfounders. Specifically, $v_{2j-1} = 1$ or 2 according to whether the grand paternal or grand maternal allele is transmitted in the paternal meiosis to the jth nonfounder. v_{2j} carries the similar information for the corresponding maternal meiosis, namely, $v_{2j} = 3$ or 4 according to whether the grand paternal or grand maternal allele was transmitted in the maternal meiosis to the jth nonfounder.

In practice, the genetic markers do not always allow us to determine the true inheritance vector. In this case, the inheritance distribution is the conditional probability distribution over the possible inheritance vectors that conform with the alleles observed at t, which we denote by $p\{v(t) = w\}$ for all inheritance vectors $w \in V$; here V is the set of all possible inheritance vectors. In the absence of any genotypic information, all inheritance vectors are equally likely according to Mendel's first law; the probability distribution is uniform.

For segregation analysis, we employed latent variables to reflect the "imaginative" genetic effects in (2.1). In linkage analysis, we have genetic markers that flow through the inheritance vector. Thus, we can still use (2.1) for linkage analysis except that $\mathbf{a}_j^i$ should be $(U_1^i, U_{2,v_{2j-1}}^i + U_{2,v_{2j}}^i)$. On one hand, we have a reduced number of latent variables. On the other hand, many of the latent variables depend on each other through the inheritance vectors. The computation of the likelihood would be summed over all inheritance vectors w in V, in addition to the probability space of the remaining independent latent variables. Because of this connection and distinction, the challenges in the linkage analysis of ordinal traits are, to a great extent, similar to those in segregation analysis of ordinal traits, for example, the asymptotic mixture of χ^2-distributions and the need to introduce the penalized likelihood (Liang and Rathouz, 1999; Zhang, Feng and Zhu, 2003).

2.1.4 *Association test*. As discussed above, linkage analysis focuses on testing the position of a marker, although it has been difficult to replicate findings in linkage studies of mental disorders. An association analysis, however, tests whether a genetic variant, including particular allele or genotype of a marker and a haplotype in several markers, is associated with a trait. Some study cohorts recruited for linkage studies have been re-genotyped for genomewide association analyses. For binary or quantitative traits, many methods have been developed and implemented. Two commonly used programs are PLINK (Purcell et al., 2007) and FBAT (Rabinowitz and Laird, 2000). To analyze an ordinal trait, Zhang, Wang and Ye (2006) introduced the following proportional odds model:

$$\text{logit}\{P(y_{ij} \le k|G_{ij})\} = \alpha_k + \beta c_{ij}, \tag{2.2}$$

where $\alpha_0, \ldots, \alpha_{K-1}$ are non-descending level parameters, β is the genetic effect. The genetic factor c_{ij} can be chosen to reflect the underlying mode of inheritance such as the number of the risk allele. Under model (2.2), the null hypothesis is $H_0 : \beta = 0$. The score statistic is

$$S = \sum_{i,j} [R^+(y_{ij}) - R^-(y_{ij})]A_{ij}, \tag{2.3}$$

where $R^+(y_{ij})$ and $R^-(y_{ij})$ are the counts of offspring in the entire sample whose trait values are greater or less than y_{ij}, respectively, and A_{ij} is the number of copies of transmitted alleles at the marker locus. Thus, Zhang, Wang and Ye (2006) proposed the following O-TDT test based on the score statistic:

$$\frac{[S - E(S|Y)]^2}{\text{Var}(S|Y)},$$

which follows a χ_1^2-distribution asymptotically. For a case-control study, $R^+(y_{ij})$ and $R^-(y_{ij})$ are the numbers of subjects whose trait values are greater or less than y_{ij}, respectively.

If we rewrite the statistic in (2.3) in a general form as $\sum_{i,j} w_{ij} A_{ij}$, this yields the classic TDT when $w_{ij} = 1$ and the QTDT (Rabinowitz, 1997) when $w_{ij} = y_{ij} - \bar{y}$, where $\bar{y}$ is the average of all y_{ij}'s. In other words, all of these tests are a weighted function of the number of transmitted alleles at the marker locus, and the choice of the weights depends on the property of the trait. With this observation, after the proper weights are computed, the existing FBAT software can be used to test the association between any trait and alleles at a marker locus.

In the following, we describe a unified method to choose weights for any kind of trait. It is straightforward to categorize a quantitative trait into any reasonable number of categories (such as deciles) and induce an ordinal scaled trait. This would allow the use of the O-TDT for a quantitative trait. In their simulation studies, Zhang, Wang and Ye (2006) demonstrated that this strategy has comparable power to the QTDT for quantitative traits. This is due to the fact that the number of categories is enough to capture most of the information in the data (e.g., following Cochran's rule; Cochran, 1977). The advantage is that the ordinal scaled test is not affected by the nonnormal distribution of a quantitative trait, and so, the unified approach is robust.

One limitation of the test proposed by Zhang, Wang and Ye (2006) is that it does not adjust for covariates. Environmental factors or covariates, such as gender and age, may confound the association of interest. In a subsequent work, Wang, Ye and Zhang (2006) generalized model (2.2) to include covariates as follows:

$$\text{(2.4)}\quad \text{logit}\{P(y_{ij} \le k | G_{ij}), z_{ij}\} = \alpha_k + \beta c_{ij} + \delta' z_{ij},$$

where z_{ij} denotes the covariates and δ is the vector of the corresponding coefficients. Consequently, the score statistic becomes

$$\text{(2.5)}\quad S = \sum_{i,j} [\hat{\gamma}(y_{ij}, z_{ij}) - \hat{\gamma}(y_{ij} - 1, z_{ij})] A_{ij},$$

where

$$\hat{\gamma}(k, z) = \frac{\exp(\hat{\alpha}_k + \hat{\delta}' z_{ij})}{1 + \exp(\hat{\alpha}_k + \hat{\delta}' z_{ij})},$$

which is the estimated probability of having a trait value no greater than k. Thus, the weight function in (2.5) is the difference between the probability of having a trait value greater than y_{ij} and the probability of having a trait value less than y_{ij}. Not surprisingly, this is in essence the same as the weight function in (2.3) where we used counts instead of frequency (or probability).

It is important to note that association analysis does not directly equate to a causal relationship. In well-designed genetic association studies, an observed association is expected to result from either a causal functional variant of a gene, or the linkage disequilibrium between the marker and a susceptibility gene. In population-based case-control studies, there are typically attempts to match cases and controls by important demographic and/or baseline information. It is not wise to over-match subjects. Alternatively, we can collect potentially important environmental variables and consider them in the association analysis. We can also use principal component analysis on the genotypes to explore whether there are "clusters" in the study cohorts that are not appropriately reflected in the environmental variables. In family-based studies, the association tends to be conditional on parental genotypes and all phenotypes.

2.1.5 *Unique challenges in analyzing ordinal traits*. Understanding the genetic mechanisms for complex diseases is challenging regardless of whether we analyze binary, ordinal or continuous traits. Any challenges that exist for analyzing binary and continuous traits remain for ordinal traits. What are the unique challenges in analyzing ordinal traits? The key difference is that there is not a simple distribution function for ordinal traits. For continuous traits, the assumption is that the traits can somehow be treated under normality, by transformation if needed. For binary traits, through a link function (e.g., logit) we only need to deal with a Bernoulli distribution. However, for ordinal traits, the two typical approaches are (a) to assume a reliability variable or a continuous latent variable or (b) to assume a proportional odds model as we presented above. The first challenge is in the estimation. The likelihood function is complicated, and based on the numerical results, it has multiple local maxima. In addition, due to identifiability (or near-identifiability), the likelihood function may be relatively flat. Combinations of the EM and other algorithms can provide practical solutions, but finding a more efficient algorithm is an open problem.

The second challenge is in the inference. When latent variables or mixture distributions are used, some of the commonly assumed regularity conditions do not hold. One solution is to use a penalized likelihood

function (Zhang, Feng and Zhu, 2003) that prevents the parameters from being near the singularity points.

Finally, model diagnostics are difficult. For example, how do we know the latent variable-based model or the proportional odds model provides an adequate fit to the data? Although the models and methods presented above do not address this and other questions, they provide a foundation for further research and improvement.

2.2 Comorbidity

The methods described above only deal with a single trait. However, comorbidity is the rule rather than the exception in studies of mental and behavioral disorders. For example, a patient may suffer from both anxiety and depression (Li and Burmeister, 2009), and the same patient may also be addicted to nicotine, alcohol, or other substances (Merikangas et al., 1998; True et al., 1999). From a data analysis perspective, we need to consider how important it is to accommodate multiple diseases/traits. In a real-data example, Chen et al. (2011) analyzed a data set from the Study of Addiction: Genetics and Environment (SAGE). By simply considering addiction to at least two of the six substances (addiction to nicotine, alcohol, marijuana, cocaine, opiates or other drugs), we were able to identify the PKNOX2 gene that reached genomewide significance level among European-origin females. Interestingly, the PKNOX2 gene has been previously identified as one of the cis-regulated genes for alcohol addiction in mice (Mulligan et al., 2006). To further delineate the benefit of considering multivariate traits, Zhu and Zhang (2009) conducted comprehensive simulation studies, considered the correlations of 0.2 and -0.2 among three quantitative traits, and demonstrated that testing correlated traits jointly is more powerful than testing a single trait at a time. Using generalized estimation equation, Lange et al. (2003) developed a family-based association test for multivariate quantitative traits (FBAT-GEE). Recently, Zhang, Liu and Wang (2010) constructed a nonparametric test based on the generalized Kendall's tau to accommodate any combination of dichotomous, ordinal, and quantitative traits.

2.2.1 *Kendall's tau.* Kendall's τ is a rank-based correlation between two variables. It contracts the probability of observing the two variables in the same order in two observations with the probability of observing the two variables in the opposite order. Specifically, for a sample of n observations $(X_1, Y_1), \ldots, (X_n, Y_n)$, two observations (X_i, Y_i) and (X_j, Y_j) are called concordant if $(X_i - X_j)(Y_i - Y_j) > 0$ and discordant if $(X_i - X_j)(Y_i - Y_j) < 0$. Then Kendall's τ is based on the difference between the numbers of concordant pairs and discordant pairs.

We introduce a kernel function,

$$\begin{aligned}&\phi((X_i, Y_i), (X_j, Y_j))\\ &\quad = \operatorname{sign}\{(X_i - X_j)(Y_i - Y_j)\}\\ &\quad = \begin{cases} 1, & \text{if } (X_i - X_j)(Y_i - Y_j) > 0,\\ -1, & \text{if } (X_i - X_j)(Y_i - Y_j) < 0,\\ 0, & \text{if } (X_i - X_j)(Y_i - Y_j) = 0,\end{cases}\end{aligned}$$

and define a U-statistic

$$U = \binom{n}{2}^{-1} \sum_{i<j} \phi((X_i, Y_i), (X_j, Y_j)). \tag{2.6}$$

Then, Kendall's τ is

$$\tau = \frac{U}{\sqrt{\operatorname{Var}_0(U)}}, \tag{2.7}$$

where $\operatorname{Var}_0(U)$ is the variance of U under the null hypothesis of no correlation between X and Y, and equal to $n(n-1)(2n+5)/18$ if X and Y are continuous variables (Hollander and Wolfe, 1999).

2.2.2 *Generalized Kendall's tau.* To test the association between genetic markers and comorbidity, Zhang, Liu and Wang (2010) generalized Kendall's tau as follows. For individuals i and j, let T_i and T_j be their vectors of traits, respectively. Then, a trait kernel is defined as

$$F_{ij} = \big(f_1\big(T_i^{(1)} - T_j^{(1)}\big), \ldots, f_p\big(T_i^{(p)} - T_j^{(p)}\big)\big)',$$

where function $f_k(\cdot)$ is the identity function for a quantitative or binary trait (Rabinowitz, 1997), or the sign function for an ordinal trait (Zhang, Wang and Ye, 2006).

Also, recall that, as in Section 2.1.4, c is the number of any chosen allele for marker genotype and let C_i refer to the C for the ith subject. Then, Zhang, Liu and Wang (2010) defined a marker kernel as

$$D_{ij} = c_i - c_j.$$

Their U-statistic is defined as

$$U = \binom{n}{2}^{-1} \sum_{i<j} D_{ij} F_{ij}. \tag{2.8}$$

The association test statistic, or generalized Kendall's tau, is $U' \operatorname{Var}_0^{-1}(U) U$, where $\operatorname{Var}_0(U)$ is the variance

of U under the null hypothesis that there is no association between marker alleles and any linked locus that influences the trait T. The test statistic follows an asymptotic χ^2-distribution under the null hypothesis.

Obviously, the statistic in (2.8) does not incorporate covariate effects. This is relatively straightforward for a single trait as was done in (2.4). Here, the traits can be a hybrid of different traits. An alternative is to impose different weights for each pair of samples in the statistic (2.8) according to the information of their covariates. The weight, denoted by $w(z_i, z_j)$ for the pair (i, j), reflects the relative importance attributed by the covariates when we derive the statistic. Zhu, Jiang and Zhang (2010) examined the following weight function. Write $z = (z^{\mathrm{co}}, z^{\mathrm{ca}})'$ with $z^{\mathrm{co}} = (z^{(1)}, \ldots, z^{(l_1)})'$ for the continuous covariates and $z^{\mathrm{ca}} = (z^{(l_1+1)}, \ldots, z^{(l)})'$ for the categorical covariates. They defined the weight function $w(z_i, z_j)$ as

$$w(z_i, z_j) = W(\|z_i^{\mathrm{co}} - z_j^{\mathrm{co}}\|) I(z_i^{\mathrm{ca}} = z_j^{\mathrm{ca}}), \tag{2.9}$$

where $W(\cdot)$ is a positive and decreasing function, for example, $W(u) = \exp(-u^2/2h^2)$, and $I(\cdot)$ is the indicator function. Then a weighted test statistic is given by

$$S = \binom{n}{2}^{-1} \sum_{i<j} D_{ij} F_{ij} w(z_i, z_j). \tag{2.10}$$

Zhang, Liu and Wang (2010) and Zhu, Jiang and Zhang (2010) showed that under the null hypothesis, the test statistic S (weighted or not) has the following asymptotic distribution conditional on all phenotypes and parental genotypes:

$$\mathrm{Var}_0^{-1/2}(S)[S - E_0(S)] \xrightarrow{d} N(0, I_p),$$

where

$$E_0(S) = \frac{2}{n-1} \sum_{i=1}^{n} \bar{u}_i E_0(C_i | M_i^{\mathrm{pa}}),$$

$$\mathrm{Var}_0(S) = \frac{4}{(n-1)^2} \cdot \sum_{i=1}^{n} \sum_{j=1}^{n} \bar{u}_i \bar{u}_j' \mathrm{Cov}_0(C_i, C_j | M_i^{\mathrm{pa}}, M_j^{\mathrm{pa}}).$$

Consequently, the following test statistic

$$\chi^2_{\mathrm{tau}} = [S - E_0(S)]' \mathrm{Var}_0^{-1}(S)[S - E_0(S)]$$

converges to χ_p^2 in distribution under the null hypothesis provided that $\mathrm{Var}_0(S)$ is full rank. In a case-control study, we do not have the markers from parents and hence the conditional expectations are replaced with the unconditional ones. Thus, the key difference in the test statistics between family studies and population studies lies in the conditioning on the parental markers. The conditioning on the parental markers gives the family studies a major advantage in removing the effect of population admixture, but family studies tend to be more difficult and expensive to carry out.

Under the alternative hypothesis, the test statistic χ^2_{tau} can be written as a weighted sum of noncentral $\chi_1^2 = \sum_{i=1}^{p} e_i \chi_1^2(\phi_i)$, where $e_1 \geq \cdots \geq e_p$ are the nonnegative eigenvalues of $\boldsymbol{\Sigma}_1^{1/2} \boldsymbol{\Sigma}_0^{-1} \boldsymbol{\Sigma}_1^{1/2}$. $\phi_i = \mu_{R_i}^2$ and μ_{R_i} is the ith component of $\boldsymbol{\mu}_R = Q \boldsymbol{\Sigma}_1^{-1/2} \boldsymbol{\mu}$, where Q is an orthonormal matrix such that $Q \boldsymbol{\Sigma}_1^{1/2} \boldsymbol{\Sigma}_0^{-1} \boldsymbol{\Sigma}_1^{1/2} Q' = \mathrm{diag}(e_1, \ldots, e_p)$. $\boldsymbol{\mu}$ is the difference in the means of S under the alternative and null hypotheses. Using the approximation theory of Pearson (1959), Solomon and Stephens (1977) and Liu, Tang and Zhang (2009), we can find a certain degree of freedom l and noncentral parametric υ such that the distribution of χ^2_{tau} can be closely approximated by $\chi_l^2(\upsilon)$. Through simulation studies, Zhu et al. confirmed that this approximation is accurate enough for power calculation.

It is noteworthy that the weight function in (2.9) is restrictive with respect to categorical covariates, especially so for ordinal covariates. The use of genomic propensity score can give rise to an alternative weight function. Specifically, for a di-allelic marker G (e.g., SNP), the genomic propensity score is the conditional probability $p_g(z) = P(G = g | Z = z)$. This probability can be fitted by a logistic regression model or proportional odds model depending on whether G is chosen as an allele type or genotype. In the latter choice, the model also depends on the mode of inheritance. In the current genomewide association studies, we usually only have genotypes and cannot distinguish the phases of individual alleles. Thus, we have to construct genomic propensity scores by considering various modes of inheritance. Once the genomic propensity score is estimated, it can be treated as a numerical covariate and then we can use (2.9) again.

2.2.3 *Examples*. Zhang, Liu and Wang (2010) reanalyzed a data set from the Collaborative Study on the Genetics of Alcoholism (COGA) (Begleiter, 1995; Edenberg et al., 2005). The data came from a multicenter (9 sites) consortium that recruited study participants by requiring every proband to meet two alcohol dependence diagnostic criteria based on DSM-IV-R (American Psychiatric Association, 1994). The first-degree relatives of the probands were invited into the

study. Zhang, Liu and Wang (2010) included a total of 1614 individuals from 143 families. They considered three phenotypes: (1) alcohol DX-DSM3R + Feighner; (2) maximum number of drinks in a 24-hour period; and (3) the response to "spent so much time drinking, had little time for anything else." Using the first phenotype alone, the p-value of the association between a peak marker D7S679 on chromosome 7 and the trait was 0.0019. However, when the three traits are analyzed together, D7S679 remains the peak marker, and the p-value is reduced to 0.00055, demonstrating the possibility that the other two phenotypes enhanced the association signal. If the other two phenotypes are analyzed alone, the analysis did not lead to anything worthy of further attention.

In the analysis cited above, the association was assessed without considering covariates. In a follow-up analysis, Zhu, Jiang and Zhang (2010) considered two important covariates: age at interview and sex. When these two covariates were controlled for, the p-value of the association between the peak marker D7S679 and the three phenotypes went down further to 0.000313.

3. DISCUSSION

Studying comorbidity is a significant issue in mental and behavioral research, dating back to a century ago (Cannon and Rosanoff, 1911). This is challenging due to a lack of statistical methods that accommodate the complexity of comorbidity. While dealing with comorbidity in genetic studies is the focus of this review, it is achieved through gradual development, and accumulation of methods. Various challenges are dealt with along the way.

Although I focused on the analysis of ordinal traits and applications in mental health, the presented methods are closely related to robust and rank-based methods for binary and quantitative traits. Furthermore, ordinal traits arise in studies of diseases besides mental illnesses, such as cancer (specifically, different stages).

From the statistical perspective, the methods that are presented here have broad applications beyond genetic association studies. From college admissions, to job searches, to scientific investigations, we make inferences based on multidimensional data. It is important and imperative to consider and develop inferential tools for multivariate outcomes, particularly when the outcomes are discrete. There is extensive literature on the statistical analysis of multivariate normal variables as well as on nonparametric tests for a single variable of nonnormal distribution. However, few options are available for the inference when we have multiple nonnormally distributed variables and potential hybrids of continuous and discrete variables. To overcome this challenge, I presented several useful statistical techniques such as the rank-based U-statistics and the kernel-based weighted statistics to accommodate the mix of continuous and discrete outcomes and the presence of important covariates.

ACKNOWLEDGMENTS

This work is supported in part by National Institute on Drug Abuse Grant R01DA016750. The author wishes to thank Professor David Madigan for his encouragement on this review. He also thanks Dr. Gang Zheng and Professor Joseph Gastwirth for helpful discussions and comments, and Jennnifer Brennan for careful reading and comments.

REFERENCES

ABELSON, J. F., KWAN, K. Y., O'ROAK, B. J., BAEK, D. Y., STILLMAN, A. A., MORGAN, T. M., MATHEWS, C. A., PAULS, D. L., RASIN, M.-R., GUNEL, M., DAVIS, N. R., ERCAN-SENCICEK, A. G., GUEZ, D. H., SPERTUS, J. A., LECKMAN, J. F., LEON S. DURE, T., KURLAN, R., SINGER, H. S., GILBERT, D. L., FARHI, A., LOUVI, A., LIFTON, R. P., SESTAN, N. and STATE, M. W. (2005). Sequence variants in SLITRK1 are associated with Tourette's syndrome. *Science* **310** 317–320.

ABREU, P. C., GREENBERG, D. A. and HODGE, S. E. (1999). Direct power comparisons between simple lod scores and NPL scores for linkage analysis in complex diseases. *Am. J. Hum. Genet*. **65** 847–857.

ALLISON, D. B. (1997). Transmission-disequilibrium tests for quantitative traits. *Am. J. Hum. Genet*. **60** 676–690.

ALMASY, L. and BLANGERO, J. (1998). Multipoint quantitative-trait linkage analysis in general pedigrees. *Am. J. Hum. Genet*. **62** 1198–1121.

AMERICAN PSYCHIATRIC ASSOCIATION (1994). *Diagnostic and Statistical Manual of Mental Disorders*, 4th ed. American Psychiatric Association Press, Washington, DC.

AMOS, C. I. (1994). Robust variance-components approach for assess-ing genetic linkage in pedigrees. *Am. J. Hum. Genet*. **54** 535–543.

ARKING, D. E., PFEUFER, A., POST, W., KAO, W. H. L., NEWTON-CHEH, C., IKEDA, M., WEST, K., KASHUK, C., AKYOL, M., PERZ, S., JALILZADEH, S., ILLIG, T., GIEGER, C., GUO, C.-Y., LARSON, M. G., WICHMANN, H. E., MARBÁN, E., O'DONNELL, C. J., HIRSCHHORN, J. N., KÄÄB, S., SPOONER, P. M., MEITINGER, T. and CHAKRAVARTI, A. (2006). A common genetic variant in the NOS1 regulator NOS1AP modulates cardiac repolarization. *Nat. Genet*. **38** 644–651.

BABIKER, A. and CUZICK, J. (1994). A simple frailty model for family studies with covariates. *Stat. Med*. **13** 1679–1692.

BEGLEITER, E. A. H. (1995). The collaborative study on the genetics of alcoholism. *Alcohol Health Res. World* **19** 228–236.

BLACKWELDER, W. C. and ELSTON, R. C. (1985). A comparison of sib-pair linkage tests for disease susceptibility loci. *Genet. Epidemiol.* **2** 85–97.

BLANGERO, J. and ALMASY, L. (1997). Multipoint oligogenic linkage analysis of quantitative traits. *Genet. Epidemiol.* **14** 959–964.

CANNINGS, C., THOMPSON, E. A. and SKOLNICK, M. H. (1978). Probability functions on complex pedigrees. *Adv. Appl. Probab.* **10** 26–61. MR0490038

CANNON, G. L. and ROSANOFF, A. J. (1911). Preliminary report of a study of heredity in insanity in the light of the Mendelian laws. Reprinted from *J. Nervous and Mental Disorders* **38** 272–279.

CARTER, C. L. and CHUNG, C. S. (1980). Segregation analysis of schizophrenia under a mixed genetic model. *Hum. Hered.* **30** 350–356.

CHEN, X., LIU, C. T., ZHANG, M. Z. and ZHANG, H. P. (2007). A forest-based approach to identifying gene and gene–gene interactions. *Proc. Natl. Acad. Sci. USA* **104** 19199–19203.

CHEN, X., CHO, K., SINGER, B. H. and ZHANG, H. P. (2011). The nuclear transcription factor PKNOX2 is a candidate gene for substance dependence in European-origin women. *PLoS ONE* **6** e16002.

COCHRAN, W. G. (1977). *Sampling Techniques*, 3rd ed. Wiley, New York. MR0474575

DEMPSTER, A. P., LAIRD, N. M. and RUBIN, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *J. Roy. Statist. Soc. Ser. B* **39** 1–38. MR0501537

DUERR, R. H., TAYLOR, K. D., BRANT, S. R., RIOUX, J. D., SILVERBERG, M. S., DALY, M. J., STEINHART, A. H., ABRAHAM, C., REGUEIRO, M. and GRIFFITHS, A. ET AL. (2006). A genomewide association study identifies IL23R as an inflammatory bowel disease gene. *Science* **314** 1461–1463.

EDENBERG, H. J., BIERUT, L. J., BOYCE, P., CAO, M., CAWLEY, S., CHILES, R. and DOHENY, K. F. (2005). Description of the data from the Collaborative Study on the Genetics of Alcoholism (COGA) and single-nucleotide polymorphism genotyping for Genetic Analysis Workshop 14. *BMC Genetics* **6** S2.

ELSTON, R. C. and STEWARD, J. (1971). A general model for the analysis of pedigree data. *Hum. Hered.* **21** 523–542.

FENG, R., LECKMAN, J. and ZHANG, H. P. (2004). Linkage analysis of ordinal traits for pedigree data. *Proc. Natl. Acad. Sci. USA* **101** 16739–16744.

GASTWIRTH, J. L. (1966). On robust procedures. *J. Amer. Statist. Assoc.* **61** 929–948. MR0205397

GASTWIRTH, J. L. (1985). The use of maximin efficiency robust tests in combining contingency tables and survival analysis. *J. Amer. Statist. Assoc.* **80** 381–384. MR0792737

GOLDGAR, D. E. (1990). Multipoint analysis of human quantitative genetic variation. *Am. J. Hum. Genet.* **47** 957–967.

GOLDSTEIN, D. B. (2009). Common genetic variation and human traits. *N. Eng. J. Med.* **360** 1696–1698.

GUO, S. W. and THOMPSON, E. A. (1992). A Monte Carlo method for combined segregation and linkage analysis. *Am. J. Hum. Genet.* **51** 1111–1126.

HEATH, A. C., TODOROV, A. A., NELSON, E. C., MADDEN, P. A. F., BUCHOLZ, K. K. and MARTIN, N. G. (2002). Gene–environment interaction effects on behavioral variation and risk of complex disorders: The example of alcoholism and other psychiatric disorders. *Twin Research* **5** 30–37.

HESTON, L. L. (1966). Psychiatric disorders in foster home reared children of schizophrenic mothers. *Bristish J. Psychiatry* **112** 819–825.

HINDORFF, L. A., SETHUPATHY, P., JUNKINSA, H. A., RAMOSA, E. M., MEHTAC, J. P., COLLINSB, F. S. and MANOLIOA, T. A. (2009). Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl. Acad. Sci. USA* **106** 9362–9367.

HOLLANDER, M. and WOLFE, D. A. (1999). *Nonparametric Statistical Methods*, 2nd ed. Wiley, New York. MR1666064

HOPPER, J. L. (1989). Modelling sibship environment in the regressive logistic model for familial disease. *Genet. Epidemiol.* **6** 235–240.

HORVATH, S. and LAIRD, N. M. (1998). A discordant-sibship test for disequilibrium and linkage: No need for parental data. *Am. J. Hum. Genet.* **63** 1886–1897.

KAO, W. H. ET AL. (2009). MYH9 is associated with nondiabetic end-stage renal disease in African-Americans. *Nat. Genet.* **40** 1185–1192.

KETY, S. S., ROSENTHAL, D. and WENDER, P. (1978). Genetic relationships within the schizophrenia spectrum: Evidence from adoption studies. In *Critical Issues in Psychiatric Diagnosis* (R. L. Spitzer and D. F. Klein, eds.) 213–223. Raven Press, New York.

KLEIN, R. J., ZEISS, C., CHEW, E. Y., TSAI, J.-Y., SACKLER, R. S., HAYNES, C., HENNING, A. K., SANGIOVANNI, J. P., MANE, S. M., MAYNE, S. T., BRACKEN, M. B., FERRIS, F. L., OTT, J., BARNSTABLE, C. and HOH, J. (2005). Complement factor H polymorphism in age-related macular degeneration. *Science* **308** 385–389.

KNAPP, M. (1999). Using exact *P* values to compare the power between the reconstruction-combined transmission/disequilibrium test and the sib transmission/disequilibrium test. *Am. J. Hum. Genet.* **65** 1208–1210.

KOPP, J. B., SMITH, M. W., NELSON, G. W., JOHNSON, R. C., FREEDMAN, B. I., BOWDEN, D. W., OLEKSYK, T., MCKENZIE, L. M., KAJIYAMA, H., AHUJA, T. S., BERNS, J. S., BRIGGS, W., CHO, M. E., DART, R. A., KIMMEL, P. L., KORBET, S. M., MICHEL, D. M., MOKRZYCKI, M. H., SCHELLING, J. R., SIMON, E., TRACHTMAN, H., VLAHOV, D. and WINKLER, C. A. (2008). MYH9 is a major-effect risk gene for focal segmental glomerulosclerosis. *Nat. Genet.* **40** 1175–1184.

KRAEPELIN, E. (1899). *Psychiatrie*: *Ein Lehrbuch fur Studirende und Aerzte*. Barth, Leipzig.

KRUGLYAK, L., DALY, M. J., REEVE-DALY, M. P. and LANDER, E. S. (1996). Parametric and nonparametric linkage analysis: A unified multipoint approach. *Am. J. Hum. Genet.* **58** 1347–1363.

LANDER, E. S. and GREEN, P. (1987). Construction of multilocus genetic linkage maps in humans. *Proc. Natl. Acad. Sci. USA* **84** 2363–2367.

LANGE, C., SILVERMAN, E. K., XU, X., WEISS, S. T. and LAIRD, N. M. (2003). A multivariate family-based association test using generalized estimating equations: FBAT-GEE. *Biostatistics* **4** 195–306.

LI, H. Z. and THOMPSON, E. (1997). Semiparametric estimation of major gene and family-specific random effects for age of onset. *Biometrics* **53** 282–293.

LI, M. D. and BURMEISTER, M. (2009). New insights into the genetics of addiction. *Nat. Rev. Genet.* **10** 225–231.

LIANG, K.-Y. and RATHOUZ, P. J. (1999). Hypothesis testing under mixture models: Application to genetic linkage analysis. *Biometrics* **55** 65–74. MR1705673

LIU, H., TANG, Y. and ZHANG, H. H. (2009). A new chi-square approximation to the distribution of non-negative definite quadratic forms in non-central normal variables. *Comput. Statist. Data Anal.* **53** 853–856. MR2657050

LIU, Y., TRITCHLER, D. and BULL, S. B. (2002). A unified framework for transmission-disequilibrium test analysis of discrete and continuous traits. *Genet. Epidemiol.* **22** 26–40.

LUNETTA, K. L., FARONE, S. V., BIEDERMAN, J. and LAIRD, N. M. (2000). Family based tests of association and linkage that used unaffected sibs, covariates, and interactions. *Am. J. Hum. Genet.* **66** 605–614.

MARTIN, E. R., MONKS, S. A., WARREN, L. L. and KAPLAN, N. L. (2000). A test for linkage and association in general pedigrees: The pedigree disequilibrium test. *Am. J. Hum. Genet.* **67** 146–154.

MERIKANGAS, K. R., STOLAR, M., STEVENS, D. E., GOULET, J., PREISIG, M. A., FENTON, B., ZHANG, H., O'MALLEY, S. S. and ROUNSAVILLE, B. J. (1998). Familial transmission of substance use disorders. *Arch. Gen. Psychiatry* **55** 973–979.

MORTON, N. E. (1955). Sequential tests for the detection of linkage. *Am. J. Hum. Genet.* **7** 277–318.

MULLIGAN, M. K., PONOMAREV, I., HITZEMANN, R. J., BELKNAP, J. K., TABAKOFF, B., HARRIS, R. A., CRABBE, J. C., BLEDNOV, Y. A., GRAHAME, N. J., PHILLIPS, T. J., FINN, D. A., HOFFMAN, P. L., IYER, V. R., KOOB, G. F. and BERGESON, S. E. (2006). Toward understanding the genetics of alcohol drinking through transcriptome meta-analysis. *Proc. Natl. Acad. Sci. USA* **103** 6368–6373.

OTT, J. (1974). Estimation of the recombination fraction in human pedigrees: Efficient computation of the likelihood for human linkage studies. *Am. J. Hum. Genet.* **26** 588–597.

OTT, J. (1999). *Analysis of Human Genetic Linkage*, 3rd ed. Johns Hopkins Univ. Press, Baltimore, MD.

PAULS, D. L. and LECKMAN, J. F. (1986). The inheritance of Gilles de la Tourette's syndrome and associated behaviors. Evidence for autosomal dominant transmission. *New Eng. J. Med.* **315** 993–997.

PEARSON, E. S. (1959). Note on an approximation to the distribution of non-central χ^2. *Biometrika* **46** 364. MR0109380

PLOMIN, R., DEFRIES, J. C., MCCLEARN, G. E. and RUTTER, M. (1997). *Behavioral Genetics*, 3rd ed. Freeman, New York.

PURCELL, S., NEALE, B., TODD-BROWN, K., THOMAS, L., FERREIRA, M. A. R., BENDER, D., MALLER, J., SKLAR, P., DE BAKKER, P. I. W., DALY, M. J. and SHAM, P. C. (2007). PLINK: A toolset for whole-genome association and population-based linkage analysis. *Am. J. Hum. Genet.* **81** 559–575.

RABINOWITZ, D. (1997). A transmission disequilibrium test for quantitative trait loci. *Hum. Hered.* **47** 342–350.

RABINOWITZ, D. and LAIRD, N. M. (2000). A unified approach to adjusting association tests for population admixture with arbitrary pedigree structure and arbitrary missing marker information. *Hum. Hered.* **50** 211–223.

RAO, S. and XU, S. (1998). Mapping quantitative trait loci for ordered categorical traits in four-way crosses. *Heredity* **81** 214–224.

RISCH, N. R. and ZHANG, H. P. (1995). Extreme discordant sib pairs for mapping quantitative trait loci in humans. *Science* **268** 1584–1589.

ROSENTHAL, D. (1972). Three adoption studies of heredity in the schizophrenic disorders. *Internat. J. Mental Health* **1** 63–75.

SCHARF, J. M., MOORJANI, P., FAGERNESS, J., PLATKO, J. V., ILLMANN, C., GALLOWAY, B., JENIKE, E., STEWART, S. E., PAULS, D. L. and THE TOURETTE SYNDROME INTERNATIONAL CONSORTIUM FOR GENETICS (2008). Lack of association between SLITRK1var321 and Tourette syndrome in a large family-based sample. *Neurology* **70** 1495–1496.

SCHORK, N. J. (1993). Extended multipoint identity-by-descent analy-sis of human quantitative traits: Efficiency, power, and modeling con-siderations. *Am. J. Hum. Genet.* **53** 1306–1319.

SIEGMUND, K. and MCKNIGHT, B. (1998). Modeling hazard functions in families. *Genet. Epidemiol.* **15** 147–171.

SOLOMON, H. and STEPHENS, M. A. (1977). Distribution of a sum of weighted chi-square variables. *J. Amer. Statist. Assoc.* **72** 881–885.

SPIELMAN, R. S. and EWENS, W. J. (1998). A sibship rest for linkage in the presence of association: The sib transmission/disequilibrium test. *Am. J. Hum. Genet.* **62** 450–458.

SPIELMAN, R. S., MCGINNIS, R. E. and EWENS, W. J. (1993). Transmission test for linkage disequilibrium: The insulin gene region and insulin-dependent diabetes mellitus (IDDM). *Am. J. Hum. Genet.* **52** 506–516.

STEINKE, J. W., BORISH, L. and ROSENWASSER, L. J. (2003). Genetics of hypersensitivity. *J. Allergy Clin. Immunol.* **111** S495–S501.

TIENARI, P. (1991). Interaction between genetic vulnerability and family environment: The Finnish adoptive family study of schizophrenia. *Acta Psychiatr. Scand.* **84** 460–465.

TRUE, W. R., HEATH, A. C., SCHERRER, J. F., XIAN, H., LIN, N., EISEN, S. A., LYONS, M. J., GOLDBERG, J. and TSUANG, M. T. (1999). Interrelationship of genetic and environmental influences on conduct disorder and alcohol and marijuana dependence symptoms. *Am. J. Med. Genet.* **88** 391–397.

VERGNE, L., BOURGEOIS, A., MPOUDI-NGOLE, E., MOUGNUTOU, R., MBUAGBAW, J., LIEGEOIS, F., LAURENT, C., BUTEL, C., ZEKENG, L., DELAPORTE, E. and PEETERS, M. (2003). Biological and genetic characteristics of HIV infections in Cameroon reveals dual group M and O infections and a correlation between SI-inducing phenotype of the predominant CRF02_AG variant and disease stage. *Virology* **310** 254–266.

WANG, X. Q., YE, Y. Q. and ZHANG, H. P. (2006). Family-based association tests for ordinal traits adjusting for covariates. *Genet. Epidemiol.* **30** 728–736.

WEINBERG, C. R. (1999). Allowing for missing parents in genetic studies of case-parental triads. *Am. J. Hum. Genet.* **64** 1186–1193.

WHITTEMORE, A. S. (1996). Genome scanning for linkage: An overview. *Am. J. Hum. Genet.* **59** 704–716.

XU, S. and XU, C. (2006). A multivariate model for ordinal trait analysis. *Heredity* **97** 409–417.

ZHANG, H. P., FENG, R. and ZHU, H. (2003). A latent variable model of segregation analysis for ordinal traits. *J. Amer. Statist. Assoc.* **98** 1023–1034. MR2041490

ZHANG, H. P., LIU, C.-T. and WANG, X. (2010). An association test for multiple traits based on the generalized Kendall's tau. *J. Amer. Statist. Assoc.* **105** 473–481. MR2724840

ZHANG, H. P. and MERIKANGAS, K. (2000). A frailty model of segregation analysis: Understanding the familial transmission of alcoholism. *Biometrics* **56** 815–823.

ZHANG, H. P., WANG, X. Q. and YE, Y. Q. (2006). Detection of genes for ordinal traits in nuclear families and a unified approach for association studies. *Genetics* **172** 693–699.

ZHANG, M., FENG, R., CHEN, X., HU, B. and ZHANG, H. (2008). LOT: A tool for linkage analysis of ordinal traits for pedigree data. *Bioinformatics* **24** 1737–1739.

ZHENG, G., JOO, J., ZAYKIN, D., WU, C. and GELLER, N. (2009). Robust tests in genome-wide scans under incomplete linkage disequilibrium. *Statist. Sci.* **24** 503–516. MR2779340

ZHU, W. S., JIANG, Y. and ZHANG, H. P. (2010). Covariate-adjusted association tests and power calculations based on the generalized Kendall's tau. Technical report.

ZHU, W. and ZHANG, H. (2009). Why do we test multiple traits in genetic association studies? *J. Korean Statist. Soc.* **38** 1–10. MR2656857

ZHU, X., COOPER, R., KAN, D., CAO, G. and WU, X. (2005). A genome-wide linkage and association study using COGA data. *BMC Genet.* **6** S128.