RESEARCH ARTICLE

Genetic Epidemiology

OFFICIAL JOURNAL
INTERNATIONAL GENETIC EPIDEMIOLOGY SOCIETY
www.geneticepi.org

A Bivariate Mann-Whitney Approach for Unraveling Genetic Variants and Interactions Contributing to Comorbidity

Yalu Wen,[1] Daniel J. Schaid,[2] and Qing Lu[1]*

[1]*Department of Epidemiology and Biostatics, Michigan State University, East Lansing, Michigan;* [2]*Division of Biostatistics, Mayo Clinic, Rochester, Minnesota*

Received 6 July 2012; Revised 3 December 2012; accepted revised manuscript 6 December 2012.
Published online in Wiley Online Library (wileyonlinelibrary.com). DOI 10.1002/gepi.21709

ABSTRACT: **Although comorbidity among complex diseases (e.g., drug dependence syndromes) is well documented, genetic variants contributing to the comorbidity are still largely unknown. The discovery of genetic variants and their interactions contributing to comorbidity will likely shed light on underlying pathophysiological and etiological processes, and promote effective treatments for comorbid conditions. For this reason, studies to discover genetic variants that foster the development of comorbidity represent high-priority research projects, as manifested in the behavioral genetics studies now underway. The yield from these studies can be enhanced by adopting novel statistical approaches, with the capacity of considering multiple genetic variants and possible interactions. For this purpose, we propose a bivariate Mann-Whitney (BMW) approach to unravel genetic variants and interactions contributing to comorbidity, as well as those unique to each comorbid condition. Through simulations, we found BMW outperformed two commonly adopted approaches in a variety of underlying disease and comorbidity models. We further applied BMW to datasets from the *Study of Addiction: Genetics and Environment*, investigating the contribution of 184 known nicotine dependence (ND) and alcohol dependence (AD) single nucleotide polymorphisms (SNPs) to the comorbidity of ND and AD. The analysis revealed a candidate SNP from *CHRNA5*, rs16969968, associated with both ND and AD, and replicated the findings in an independent dataset with a *P*-value of 1.06×10^{-03}.**
Genet Epidemiol 00:1–8, 2013. © 2013 Wiley Periodicals, Inc.

KEY WORDS: **forward selection; high-order interaction; *CHRNA5*; nicotine dependence; alcohol dependence**

Introduction

Radical breakthroughs in biotechnologies have made it possible to rapidly and accurately genotype millions of single nucleotide polymorphisms (SNPs) at an affordable cost. Benefiting from these high-throughput technologies and the HapMap project [The International HapMap Consortium, 2003], there has been significant progress in genome-wide association studies focused on discovering novel genetic variants that contribute to complex human diseases [Barrett et al., 2008; Baum et al., 2008; Bierut et al., 2004, 2010; Caporaso et al., 2009; Ferreira et al., 2008; Moskvina et al., 2009; Schlaepfer et al., 2008b; Sklar et al., 2008; Treutlein et al., 2009; Wang et al., 2011; Wellcome Trust Case Control Consortium, 2007]. With the increase in genetic findings, converging evidence has revealed that the same genetic variants could be associated with multiple related-disease outcomes. For example, recent studies have provided evidence that the neuronal nicotinic acetylcholine receptor (*nAChRs*) subunit genes may play an important role in the common pathophysiological pathway of nicotine dependence (ND) and alcohol dependence (AD) [Clark et al., 2001; John et al., 2003; Riala et al., 2004]. Similarly, clinical and epidemiological studies have suggested a high degree of comorbidity between bipolar disorder and migraine, which could be partially explained by a shared genetic component [Bowden et al., 2000; Dilsaver et al., 2009a, b; Oedegaard et al., 2010a, b]. Despite these findings, the pathophysiology and etiology of disease comorbidity remain largely unknown [Oedegaard et al., 2010a]. It is of great importance to identify genetic variants and environmental determinants common to disease comorbidity, as well as those are unique to each condition, as this helps elucidate the causes of comorbidity, and promotes new diagnostic and therapeutic strategies for both diseases.

The concept of "comorbidity" was first introduced in the 1970s by Feinstein. It stands for the scenario in which "a distinct clinical entity" occurred together with a specific disease under study [de Groot et al., 2003; Feinstein, 1970; Maj, 2005]. Recently, multicomorbidity has been introduced, referring to a scenario where multiple medical conditions occur in one person without an emphasis on the presence of a specific disease [Bayliss et al., 2008; Valderas et al., 2009]. Both comorbidity and multicomorbidity are used in the domains of clinical care, epidemiology studies, and health service policies [Bayliss et al., 2008; Campbell-Scherer, 2010; de Groot et al., 2003; Feinstein, 1970; Gijsen et al., 2001; Maj, 2005; Valderas et al., 2009]. In the rest of this

Supporting Information is available in the online issue at wileyonlinelibrary.com.

*Correspondence to: Qing Lu, B601 West Fee Hall, Michigan State University, 909 Fee Road, East Lansing, MI 48824. E-mail: qlu@epi.msu.edu

paper, we use comorbidity to refer to both comorbidity and multicomorbidity.

The relation between comorbid conditions is complex and presents in various forms. To describe the underlying mechanisms leading to disease comorbidity, Neale and Kendler have proposed 13 theoretical comorbidity models [Neale and Kendler, 1995; Rhee et al., 2004]. The simplest scenario is that comorbid conditions are independent of each other and occur together simply by chance or due to a third distinct disease [Neale and Kendler, 1995]. Comorbidity can also be the cause or consequence of one of the comorbid conditions, with possible reciprocal causality [de Groot et al., 2003; Neale and Kendler, 1995; Simonoff, 2000]. Another common scenario is that comorbid conditions share the same or correlated risk factors, which makes the comorbid conditions more likely to occur together [Lind et al., 2010; Neale and Kendler, 1995; Youngstrom et al., 2010]. In certain circumstances, comorbidity may also be due to the fact that the comorbid conditions are alternate manifestations of a single liability [Neale and Kendler, 1995].

A common approach for studying comorbidity is the composite phenotype (COM) approach, in which the "cases" are defined as individuals with all comorbid conditions, while the "controls" are defined as individuals with none of the comorbid conditions [Bierut et al., 2004; Lind et al., 2010; Oedegaard et al., 2010a]. Though easy to implement, such an approach does not take individuals with only one of the comorbid conditions into account. As a consequence of the reduced sample size, it may lack the power to catch pathophysiological pathways underlying the disease. In addition, COM is designed to identify common genetic variants leading to comorbidity, but not unique genetic variants for each disease outcome. Another approach to study comorbidity is the EITHER approach, which defines cases as individuals with at least one of the comorbid conditions and the controls as individuals without the comorbid conditions. However, similar to COM, EITHER cannot differentiate common genetic variants from unique genetic variants. Moreover, EITHER is subject to reduced power if the genetic etiologies of the two comorbid conditions are independent of each other. To address these limitations, we propose a bivariate Mann-Whitney (BMW) approach for comorbidity study. The proposed approach utilizes the entire sample and is capable of capturing shared genetic variants and their possible interactions contributing to disease comorbidity, as well as unique genetic variants for each disease outcome. In the following sections, we first lay out the details of the BMW approach, and then evaluate the performance of the proposed approach with simulations. Finally, we apply the new approach to a large-scale dataset from the *Study of Addiction: Genetics and Environment (SAGE)* to study the comorbidity between AD and ND.

Method

Consider a comorbidity study of N unrelated individuals and G genetic markers, where we are interested in identifying shared and unique disease susceptibility markers contributing to comorbid conditions. Let Y_k ($k = 1, 2$) be the response measurement of k^{th} condition and $Z = \{Z_1, Z_2, \ldots, Z_G\}$ be the measurement of G markers, where $Y_k = 1$ and $Y_k = 0$ for individuals with and without the k^{th} condition, respectively. Comorbidity stands for the scenario where two conditions occur in the same person (i.e., $Y_1 = Y_2 = 1$). The BMW approach first applies a Mann-Whitney based forward selection algorithm [Lu et al., 2012] to search for genetic variants and interactions predisposing to each of the two conditions. The algorithm starts with a null model without any genetic markers and then gradually selects disease-susceptibility markers into the model. In step one, it searches all G genetic markers for the marker most strongly associated with the given condition. In step two, it searches for the second marker that is most related to the condition, considering its possible interaction with the marker selected at step one. The whole process continues until it reaches a full model. K-fold cross-validation is then used to choose the most parsimonious model.

By applying Mann-Whitney based forward selection to each of two conditions, we identify two sets of disease-susceptibility markers, $X_1 = \{Z_{p_1}, Z_{p_2}, \ldots, Z_{p_M}\}$, $M \leq G$ and $X_2 = \{Z_{q_1}, Z_{q_2}, \ldots\ldots, Z_{q_S}\}$, $S \leq G$, for conditions 1 and 2, respectively. Let $X_C = X_1 \cap X_2$ be the common set of markers shared by both diseases, $X_{U_1} = X_1 \cap \overline{X_C}$ be the subset of markers unique to disease 1, and $X_{U_2} = X_2 \cap \overline{X_C}$ be the subset of markers unique to disease 2. We use $X_{i,C}$ to denote the measurements of shared loci for subject i, and use X_{i,U_1} and X_{i,U_2} to denote the measurements of loci unique to disease 1 and loci unique to disease 2 for subject i, respectively. The likelihood ratio (LR) of individual i for the shared markers, measuring the risk of shared makers with the disease, can be defined as $LR(X_{i,C}) = \frac{P(X_{i,C}|Y\neq 0)}{P(X_{i,C}|Y=0)}$, where $Y = \begin{cases} 1 & \text{if } Y_1 = 1 \text{ or } Y_2 = 1 \\ 0 & \text{if } Y_1 = 0 \text{ and } Y_2 = 0 \end{cases}$ and $X_C \neq \emptyset$. Similarly, we define the LR of individual i for the unique markers as $LR(X_{i,U_k}) = \frac{P(X_{i,U_k}|X_{i,C}, Y_k=1)}{P(X_{i,U_k}|X_{i,C}, Y=0)}$, where $X_C \neq \emptyset$, $X_{U_k} \neq \emptyset$ and $k = 1, 2$. In the cases where a null set occurs (i.e., no marker has been selected), we define

$$\begin{cases} LR(X_{i,U_k}) = \dfrac{P(X_{i,U_k}|Y_k = 1)}{P(X_{i,U_k}|Y = 0)}; \ LR(X_{i,C}) = 1 & X_C = \emptyset \\ LR(X_{i,U_k}) = 1 & X_{U_k} = \emptyset \end{cases}.$$

Given the LRs for shared and unique markers, a joint LR for individual i can be defined as,

$$LR_i^M = LR(X_{i,C}) \prod_{k=1}^{2} LR(X_{i,U_k}), \tag{1}$$

Based on the joint LR, we derived a BMW statistic to assess the joint association of disease-susceptibility markers, allowing for gene-gene interactions,

$$U_{BMW} = \sum_{i=1}^{N_{Y\neq 0}} \sum_{j=1}^{N_{Y=0}} \psi(LR_i^M, LR_j^M), \tag{2}$$

where $N_{Y\neq 0}$ and $N_{Y=0}$ are the number of individuals with at least one of two diseases, and the number of individuals without either of the diseases, respectively. The kernel function ψ

equals 1 if LR_i^M is greater than LR_j^M, 0.5 if equal, and 0 if less. Hypothesis testing can then be conducted to assess the significance of the joint association,

$$z = \frac{U_{BMW} - \frac{N_{Y\neq 0} N_{Y=0}}{2}}{\sqrt{S_D + S_{\bar{D}}}}, \quad (3)$$

where $S_D = \sum_{i=1}^{N_{Y\neq 0}} (\sum_{j=1}^{N_{Y=0}} \psi(LR_i^M, LR_j^M) - \frac{U_{BMW}}{N_{Y\neq 0}})^2$ and $S_{\bar{D}} = \sum_{j=1}^{N_{Y=0}} (\sum_{i=1}^{N_{Y\neq 0}} \psi(LR_i^M, LR_j^M) - \frac{U_{BMW}}{N_{Y=0}})^2$, derived based on the result of Lu et al. [DeLong et al., 1988; Lu et al., 2012]. Under the null, Z asymptotically follows a standard normal distribution.

Simulations

Scenario I

In the first set of simulations, we compared the performance of BMW with COM and EITHER under a variety of comorbidity correlation models. We simply simulated two comorbid diseases and considered a series of comorbidity models: a model where two diseases were unrelated; a model where two diseases shared one SNP; a model where two diseases shared a two-locus interaction; and a model where two diseases were associated with the exact same disease susceptibility loci. Each comorbid disease was associated with a two-locus interaction and an independent SNP, where we assumed the two-locus interaction followed a multiplicative-interaction model or a threshold-interaction model [Marchini et al., 2005], and the independent SNP was additive. In the multiplicative-interaction model, the odds increases multiplicatively with the number of disease-susceptibility alleles given both loci have at least one disease-susceptibility allele. In the threshold model-interaction, we can group two-locus genotypes into two risk groups: (1) a single high-risk group for all individuals having at least one of the disease-susceptibility alleles at each of the two loci, and (2) a common low-risk group for all other individuals [Marchini et al., 2005]. For the multiplicative-interaction model, we assume the odds ratios of the interaction loci and the independent loci to be 1.45. For the threshold-interaction model, the odds ratios for two interaction loci and two independent loci were assumed to be 1.7, 1.7, 1.7, and 1.65, respectively. The details of model settings were summarized in the Supporting Information (Table S1). All genetic variants were simulated under the Hardy-Weinberg Equilibrium assumption with minor allele frequencies ranging from 0.3 to 0.4. In addition to disease-susceptibility loci, we also introduced five nondisease SNPs for each disease, and randomly assigned their minor allele frequencies from a uniform distribution ranging from 0.1 to 0.5. For each underlying correlation model, 1,000 replicates were simulated, each comprised of 1,000 control individuals and 1,000 affected individuals with at least one of the comorbid conditions. We analyzed each replicate by using the proposed BMW approach, the COM approach, and the EITHER approach. To be consistent with BMW, the same Mann-Whitney based forward selection algorithm was used in COM and EITHER to search for genetic variants and interactions predisposing to the comorbidity of diseases. However, unlike BMW, COM, and EITHER consider the disease outcome as a univariate variable, and thus use a different test statistic for association test [Lu et al., 2012]. The difference between COM and EITHER is how they define cases and controls. COM considers individuals with both comorbid conditions as cases, while EITHER treats individuals with at least one of the comorbid conditions as cases. In a study where the disease-associated genetic variants and interactions to be tested are predetermined (e.g., in a replication study), asymptotic test can be used to assess the significance of the association. However, in a study where the disease model is unknown (e.g., in an initial study), selecting the model and performing the asymptotic test on the same dataset could affect the null distribution, leading to an inflated Type 1 error. Therefore, a permutation process, where the phenotype was randomly permuted, was implemented in the simulation to generate the empirical null distribution. In the permutation process, we performed 1,000 permutation to form the empirical distribution, and obtained the empirical P-value by comparing the observed statistic to the empirical null distribution.

The Type I error and power for each comorbidity model were summarized in Table 1. The results showed that the Type I errors from all three approaches were well controlled at the level of 0.05. We also observed that the power of COM increased with the increase of shared genetic components. In an extreme case, when the two comorbid conditions shared the same genetic loci, the power of COM attained its highest value, which can be largely explained by the increasing number of individuals with both comorbid conditions. Nevertheless, when the two comorbid conditions were independent and the simultaneous manifestation of both diseases occurred only by chance, the power of COM was significantly reduced. The performance of EITHER also highly depended on the underlying disease models. EITHER attained high power if two diseases shared the same or similar disease mechanisms, and had low power if two diseases were independent of each other. Compared with COM and EITHER, BMW attained higher, or at least equivalent, power under all models. The performance of BMW was also less affected by the relationship between comorbid conditions, remaining almost the same across all models. While we expect that COM has no power under the model where two diseases are independent with no shared loci, the results showed that COM obtained power of 0.530 and 0.561 under the multiplicative-interaction model and the threshold-interaction model, respectively. As we demonstrated in a later simulation (Scenario III), the power of COM can also be partially explained by loci unique to each condition (i.e., if a locus is strongly associated with one of the comorbid conditions, it could also have an effect on a subset of individuals with two conditions). However, the drawback of both COM and EITHER is that, unlike BMW, they cannot distinguish the shared and unique disease-susceptibility loci.

Table 1. Type I error and power of BMW, COM, and EITHER under multiplicative-interaction model and threshold-interaction model

Models		Multiplicative interaction			Threshold interaction		
		BMW[a]	COM[b]	EITHER[c]	BMW[a]	COM[b]	EITHER[c]
Without shared loci							
Disease 1: A+B*C	Power	0.970	0.530	0.420	0.880	0.561	0.351
Disease 2: D+E*F	Type I error	0.040	0.050	0.020	0.041	0.050	0.010
Shared one locus							
Disease 1: ***A***+B*C	Power	0.930	0.650	0.760	0.931	0.570	0.730
Disease 2: ***A***+D*E	Type I error	0.055	0.010	0.043	0.049	0.041	0.041
Shared two loci							
Disease 1: A+***B*C***	Power	0.969	0.826	0.967	0.950	0.680	0.900
Disease 2: D+***B*C***	Type I error	0.049	0.050	0.038	0.058	0.060	0.068
Shared all loci							
Disease 1: ***A+B*C***	Power	0.989	0.929	0.997	0.989	0.809	0.990
Disease 2: ***A+B*C***	Type I error	0.055	0.041	0.035	0.054	0.031	0.012

Bold and Italic letters represent the shared loci between two diseases.
[a] Bivariate Mann-Whitney (BMW) approach.
[b] Composite phenotype approach, in which cases were defined as individuals with both diseases and the controls were healthy individuals.
[c] The cases were defined as individuals with at least one of the comorbid diseases, and the controls were healthy individuals.

Scenario II

In this set of simulations, we varied both underlying disease models and relations between two comorbid conditions, and evaluated their impact on the three approaches. We started with a simple model with a two-locus interaction and two independent loci, and then considered a more complex model involving a high-order interaction (i.e., a three-locus interaction) and a model involving more than one interaction (i.e., two two-locus interactions). The common disease-susceptibility loci contributing to both diseases were assumed to be (1) the interacting loci, (2) the interacting loci and one independent locus, and (3) two independent loci. Two types of interaction models, a multiplicative-interaction model and a threshold-interaction model [Marchini et al., 2005], were considered in the simulation. The odds ratios for the multiplicative-interaction model were set within the range of 1.3 to 1.4, while the odds ratios for the threshold-interaction model were set between 1.3 and 1.6. The details of the simulation settings were summarized in the Supporting Information (Table S2).

The Type I errors were well controlled at the level of 0.05 for all three approaches (Table 2). Similar to the result from simulation I, the performance of COM and EITHER was highly dependent on underlying disease models. Compared with COM and EITHER, BMW was robust to a variety of relations between two comorbid conditions, and attained higher power under all of simulated disease models, regardless of the complexity of the disease models and the different types of interaction models.

Scenario III

One of the unique features of BMW is that it can distinguish unique loci predisposing each comorbid condition from common loci contributing to comorbidity. To demonstrate this feature, a simple disease model was simulated where each of the two comorbid diseases was associated with a common two-locus interaction and a unique locus. We varied the ratio of the effect size of the two-locus interaction to that of the independent loci, and calculated the probability of misclassifying a unique locus as a shared locus. In addition, for BMW approach, we also calculated the probability of misclassifying a shared locus as a unique locus. Both multiplicative-interaction and threshold-interaction models were considered in the simulation. As EITHER is unable to identify common loci contributing to comorbidity, we only compared the performance between COM and BMW in the simulation. The details of the model settings and the results were summarized in Table 3.

As shown in Table 3, when the effect size of risk loci unique to each disease increases, the COM approach is more likely to misclassify them as common risk loci. COM considers all of the selected loci as common loci without differentiating between unique and shared loci. In contrast to COM, BMW only considers loci selected for both conditions as shared loci, while treats the remaining loci as unique loci. Thus, it has the capacity to differentiate unique and shared loci. As seen in Table 3, BMW remains a low and stable misclassification rate, regardless of the effect size of the unique loci. In addition, we also calculated the rates of misclassifying common loci as unique loci by BMW, and on average the rates were 11.2% and 12.9% for multiplicative-interaction model and threshold-interaction model, respectively.

Results

Application to ND and AD

We applied the proposed approach to datasets from *SAGE,* investigating genetic variants and interactions contributing to comorbidity between AD and ND (dbGaP study accession phs000092.v1.p1). *SAGE* is one of the largest and most comprehensive genetic datasets for addiction genetic epidemiology research [Bierut et al., 2010]. The participants of the *SAGE* were unrelated individuals selected from three independent studies: the *Collaborative Study on the Genetics*

Table 2. Type I error and power of BMW, COM, and EITHER under multiplicative-interaction model and threshold-interaction model

Models		Multiplicative interaction			Threshold interaction		
		BMW[a]	COM[b]	EITHER[c]	BMW[a]	COM[b]	EITHER[c]
Two-locus interaction models							
Disease 1: ***A***B*+*C*+D	Power	0.991	0.840	0.990	0.887	0.828	0.828
Disease 2: ***A***B*+*C*+E	Type I error	0.050	0.051	0.051	0.047	0.066	0.039
Disease 1: ***A***B*+C+D	Power	0.972	0.649	0.951	0.887	0.588	0.734
Disease 2: ***A***B*+E+F	Type I error	0.043	0.049	0.040	0.065	0.048	0.065
Disease 1: A*B+***C*+*D***	Power	0.928	0.407	0.755	0.820	0.573	0.794
Disease 2: E*F+***C*+*D***	Type I error	0.052	0.053	0.045	0.040	0.054	0.042
Three-locus interaction models							
Disease 1: ***A***B*C*+*D***+E	Power	0.931	0.793	0.905	0.928	0.847	0.880
Disease 2: ***A***B*C*+*D***+F	Type I error	0.053	0.045	0.049	0.050	0.066	0.051
Disease 1: ***A***B*C*+D+E	Power	0.900	0.729	0.779	0.901	0.820	0.829
Disease 2: ***A***B*C*+F+G	Type I error	0.045	0.045	0.048	0.039	0.053	0.049
Disease 1: A*B*C+***D*+*E***	Power	0.822	0.359	0.461	0.828	0.486	0.602
Disease 2: F*G*H+***D*+*E***	Type I error	0.053	0.040	0.046	0.041	0.069	0.041
Two two-locus interaction models							
Disease 1: ***A***B*+*C*D***+E	Power	0.996	0.876	0.984	0.975	0.964	0.940
Disease 2: ***A***B*+*C*D***+F	Type I error	0.044	0.067	0.067	0.048	0.059	0.048
Disease 1: ***A***B*+C*D+***E***	Power	0.996	0.874	0.985	0.946	0.885	0.927
Disease 2: ***A***B*+F*G +***E***	Type I error	0.054	0.048	0.061	0.050	0.060	0.050
Disease 1: A*B+C*D+***E***	Power	0.907	0.427	0.596	0.885	0.713	0.748
Disease 2: F*G+H*I+***E***	Type I error	0.039	0.064	0.053	0.052	0.043	0.049

Bold and Italic letters represent the shared loci between two diseases.
[a] Bivariate Mann-Whitney (BMW) approach.
[b] Composite phenotype approach, in which cases were defined as individuals with both diseases and the controls were defined as individuals free of diseases.
[c] The cases were defined as individuals with at least one of the comorbid disease, and the controls were defined as individuals without any co-morbid diseases.

Table 3. Misclassification rates of unique loci to common loci and misclassification rates of common loci to unique loci

Disease model	Odds ratio	Multiplicative interaction			Threshold interaction		
		$MR_{U\to C}$[a]	$MR_{C\to U}$[b]		$MR_{U\to C}$[a]	$MR_{C\to U}$[b]	
		BMW[c]	BMW[c]	COM[d]	BMW[c]	BMW[c]	COM[d]
Disease 1: ***A***B*+C	1.4;[e] 1.9;[f] 1.9[g]	0.136	0.115	0.749	0.200	0.146	0.903
	1.5;[g] 1.4;[e] 1.8;[f] 1.8[g]	0.131	0.106	0.679	0.205	0.151	0.835
	1.4;[e] 1.7;[f] 1.7;[g]	0.138	0.105	0.587	0.215	0.147	0.826
	1.4;[e] 1.6;[f] 1.6[g]	0.145	0.089	0.500	0.142	0.116	0.808
	1.4;[e] 1.5[f]	0.137	0.090	0.414	0.133	0.141	0.713
Disease 2: ***A***B*+D	1.4;[e] 1.4;[f] 1.4[g]	0.125	0.079	0.319	0.108	0.160	0.659
	1.4;[e] 1.3;[f] 1.3[g]	0.098	0.075	0.333	0.093	0.158	0.567
	1.4;[e] 1.2;[f] 1.2[g]	0.074	0.083	0.256	0.045	0.149	0.494
	1.4;[e] 1.1;[f] 1.1[g]	0.030	0.112	0.261	0.024	0.148	0.500

Bold and Italic letters represent the shared loci between two diseases.
[a] Misclassification rate of common loci to unique loci.
[b] Misclassification rate of unique loci to common loci.
[c] Bivariate Mann-Whitney approach.
[d] Composite phenotype approach, in which cases were defined as individuals with both diseases and the controls were healthy individuals.
[e] Odds ratio for the common risk loci.
[f] Odds ratio for the risk locus unique to disease 1.
[g] Odds ratio for the risk locus unique to disease 2.

of Alcoholism (COGA), the *Family Study of Cocaine Dependence (FSCD)*, and the *Collaborative Genetic Study of Nicotine Dependence (COGEND)*. The *SAGE* included standardized diagnostic assessments of DSM-IV ND and AD, for which an ample number of ND cases (n = 1,848), AD cases (n = 1,938), and controls (n = 1,590) exist (Table S3). *SAGE* genotyping is based on the Illumina Human 1M DNA Analysis BeadChip.

Based on the existing literature, we identified 152 SNPs and 32 SNPs that had been reported for potential association with ND and AD, respectively. The SNPs included in the study were those having prior association evidence with ND and AD, or those allocated in the region where significant haplotype blocks have been reported. The detailed information of these SNPs was listed in the Supporting Information (Table S4). Among these SNPs, genotypes for 127 SNPs were available in the *SAGE* dataset, and the remaining SNPs were imputed using IMPUTE2 software (IMPUTE2 version 2.2.2) [Howie et al., 2009, 2011]. Depending on whether a SNP can be imputed using HapMap reference panels, the founders of CEU and YRI from HapMap phase III or from the 1000 Genome Project, were used as the reference panels to impute genotypes of the Caucasian and African American subjects, respectively.

Table 4. Summary of models identified in the Caucasian and African American samples

Model	Sample used	*P*-values			Selected SNPs	Allele	Chromosome	Position	Gene	Disease
		COGA	FSCD	COGEND						
1	Caucasian	8.23×10^{-03}	1.06×10^{-03}	0.409	rs16969968	A/G	15	78882925	*CHRNA5*	AD, ND
2	African American	3.51×10^{-07}	0.443	0.821	rs2964911	C/T	5	163724281	*LARP4*	AD, ND
					rs1782134	C/T	14	41715568		AD
					rs10889635	A/G	1	67075575	*SGIP1*	ND

Table 5. Odds ratio for rs16969968 associated with comorbidity in the Caucasian samples

	Odds ratios		
	COGA	FSCD	COGEND
Alcohol dependence[a]	0.676 (0.511, 0.894)[b]	0.560 (0.398, 0.788)	1.148 (0.873, 1.508)
Nicotine dependence[a]	0.697 (0.518, 0.938)	0.623 (0.438, 0.885)	1.122 (0.880, 1.430)
Comorbidity[a]	0.678 (0.501, 0.919)	0.610 (0.421, 0.883)	1.202 (0.891, 1.621)

[a] GG is the reference group.
[b] 95% confidence interval.

We initiated a bivariate analysis of comorbid conditions among 184 known AD and ND SNPs by applying BMW to each of the African American and the Caucasian samples in the COGA dataset. The initial findings identified from COGA were then validated in the FSCD and COGEND datasets. In the Caucasian sample, rs16969968 (*CHRNA5*) was identified to be associated with both AD and ND, with *P*-values of 8.23×10^{-03} and 1.06×10^{-03} in COGA and FSCD, respectively. However, the findings cannot be replicated in the Caucasian sample from the COGEND dataset (*P*-value = 0.409). The lack of association between the identified SNP and the comorbidity in COGEND may partially due to the fact that nondependent smoking individuals were recruited as controls in COGEND. The effects of the genetic variants leading to comorbidity may thus be attenuated compared with the studies where nonsmoking individuals serve as controls. The lack of replication could also be due to the complexity of the trait. For instance, the ND cases were defined by DSM-IV, which was an arguably poor diagnosis criterion compared with the Fagerstrom dependence criteria. Further analysis using logistic regression was conducted to explore the association between rs16969968 and the two comorbid conditions in the Caucasian samples. Based on the logistic regression analysis, we found Caucasian individuals carrying GG genotype of rs16969968 had increased risk of having both AD and ND in COGA and FSCD, whereas the risk associated with three genotypes remained the same for the Caucasian sample in COGEND (Table 5).

In the African American sample, BMW identified a three-locus joint association model, reached an uncorrected *P*-value of 3.51×10^{-07} in COGA (Table 4). However, the finding was not replicated in the remaining two datasets (i.e., *P*-values in FSCD and COGEND were 0.443 and 0.821, respectively). The lack of association of the identified model in the African American subjects may partially due to the relatively small number of the African American in the dataset (Table S3). With a moderate number of candidate SNPs and a small sample size, the model identified in the African American sample could be just a chance finding.

Similar to the analyses using BMW, we conducted stratified analyses by applying COM and EITHER to each of the Caucasian and the African American samples. In the Caucasian samples, both COM and EITHER identified the same risk SNP, rs16969968, as the one selected by BWM. In Caucasian samples from COGA, the selected model reached *P*-values of 0.012 and 8.23×10^{-03} for COM and EITHER, respectively. Similar to the findings of BMW, the association was replicated in FSCD with *P*-values of 8.66×10^{-03} and 1.056×10^{-03} for COM and EITEHR, respectively. In the African American sample, the COM approach identified only one risk SNP, rs2964911, which is among the shared loci identified by BMW. The SNPs identified by the EITHER approach were the same as those selected by BMW. The *P*-values of the models selected by COM and EITHER in the COGA African American sample attained 9.26×10^{-03} and 3.28×10^{-04}, respectively. However, similar to BMW, the models cannot be replicated in the other datasets.

Discussion

Comorbidity among complex human diseases is believed to be caused by interplay between multiple genetic variants and environmental determinants. Identifying genetic and environmental risk predictors contributing to comorbidity will promote a better understanding of disease etiology, and may eventually lead to new diagnostic and therapeutic strategies [de Groot et al., 2003; Feinstein, 1970; Maj 2005]. The yield from the discovery process can be enhanced by adopting novel statistical approaches. A multivariate joint association approach allowing for gene-gene interactions can facilitate the detection of genetic variants and gene-gene interactions contributing to comorbid conditions. For this purpose, we developed a BMW approach for the identification of genetic variants contributing to comorbid conditions, with the consideration of high-order interactions. Similar to other Mann-Whitney based methods [Lu et al., 2012], it is a nonparametric approach, which does not assume a model of inheritance, and is free of the issues of an increasing number of parameters. BMW adopts a forward selection algorithm, which substantially reduces the searching space of interaction combinations and allows for high-order interactions. These features make BMW more appealing for a comorbidity study of complex diseases with the

consideration of possible interactions. Though the approach is illustrated with two disease outcomes, it can also be easily extended to multiple disease outcomes.

Through simulations, we have shown that BMW attained higher power than both COM and EITHER under a variety of disease models, and was more robust under different correlation models between comorbid conditions. We consider this important, as our knowledge of disease comorbidity is limited, and the underlying correlation among comorbid diseases could vary from case to case. While the COM only identifies risk variants predisposing to both diseases and the EITHER cannot differentiate the shared loci from the unique loci, BMW allows for the identification of genetic risk variants common to comorbid conditions, as well as those unique to each comorbid condition. Compared with COM, BMW makes use of the entire sample, which potentially increases the power to identify genetic variants associated with comorbid conditions, especially when the comorbidity rate is low or when there are few comorbidity individuals in the data. In an extreme case, where each dataset is designed to study one of the comorbid conditions and the information regarding the other disease statuses is not measured, COM is not applicable, as there is no affect individual according to its definition. Nevertheless, we can still use BMW in such case, as it selects risk loci for each disease and then builds an overall test to assess the association.

Twin studies have suggested a substantial genetic correlation between ND and AD [True et al., 1999]. Although the comorbidity of ND and AD is well documented, genetic variants and gene-gene interactions contributing to the comorbidity are still largely unknown[Schlaepfer et al., 2008b]. In our analysis, we identified a *CHRNA5* SNP associated with both AD and ND in the Caucasians sample of the COGA dataset, and then confirmed the finding in the FSCD dataset. Further analysis suggested Caucasian individuals carrying GG genotype of rs16969968 had increased risk of having both AD and ND. The identified SNP, rs16969968, which has been reported to be associated with ND [Berrettini et al., 2008; Saccone et al., 2007; Stevens et al., 2008; Thorgeirsson et al., 2008], is located within *CHRNA5*, a subunit gene of *nAChRs* [Berrettini et al., 2008; Bierut et al., 2007; Caporaso et al., 2009; Ehringer et al., 2007; Grucza et al., 2008; Schuckit et al., 2008; Spitz et al., 2008; Stevens et al., 2008; Thorgeirsson et al., 2008; Zeiger et al., 2008]. Although rs16969968 (*CHRNA5*) itself has not been reported to be associated with AD, the SNP rs1051730, which is in high-linkage disequilibrium with rs16969968 (*CHRNA5*; European: $r^2 = 0.902$, Japanese/Chinese: $r^2 = 1.000$) [Ware et al., 2011], had been reported to be associated with AD [Wang et al., 2009]. Previous evidence from pharmacological, epidemiological, and neurochemical studies have suggested that subunits of *nAChRs* may be a common action site for AD and ND [Aistrup et al., 1999; Butt et al., 2004; Hoft et al., 2009; Larsson and Engel, 2004; Schlaepfer et al., 2008a; Wang et al., 2009]. This finding further confirmed the important role of *nAChRs* in the common biology pathway of ND and AD. Although the association rs16969968 (*CHRNA5*) reached a statistically significant level and can be replicated in another independent dataset, follow-up studies would be needed to further replicate and study the role of *CHRNA5* in the comorbidity of ND and AD.

Acknowledgments

We thank Jim Anthony and Naomi Breslau for helpful discussions. We also appreciate critical input from the editor and two anonymous reviewers. This work was supported by the National Institute of Dental & Craniofacial Research under Award Number R03DE022379 (QL) and the U.S. Public Health Service, National Institutes of Health (NIH) under contract grant number GM065450 (DJS). Funding support for the *Study of Addiction: Genetics and Environment* (SAGE) was provided through the NIH Genes, Environment and Health Initiative [GEI] (U01 HG004422). The datasets used for the analyses described in this manuscript were obtained from dbGaP at http://www.ncbi.nlm.nih.gov/projects/gap/cgibin/study.cgi?study_id=phs000092.v1.p1 through dbGaP accession number phs000092.v1.p1.

References

Aistrup GL, Marszalec W, Narahashi T. 1999. Ethanol modulation of nicotinic acetylcholine receptor currents in cultured cortical neurons. *Mol Pharmacol* 55(1): 39–49.

Barrett JC, Hansoul S, Nicolae DL, Cho JH, Duerr RH, Rioux JD, Brant SR, Silverberg MS, Taylor KD, Barmada MM and others. 2008. Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. *Nat Genet* 40(8):955–962.

Baum AE, Hamshere M, Green E, Cichon S, Rietschel M, Noethen MM, Craddock N, McMahon FJ. 2008. Meta-analysis of two genome-wide association studies of bipolar disorder reveals important points of agreement. *Mol Psychiatry* 13(5):466–467.

Bayliss EA, Edwards AE, Steiner JF, Main DS. 2008. Processes of care desired by elderly patients with multimorbidities. *Fam Pract* 25(4):287–293.

Berrettini W, Yuan X, Tozzi F, Song K, Francks C, Chilcoat H, Waterworth D, Muglia P, Mooser V. 2008. Alpha-5/alpha-3 nicotinic receptor subunit alleles increase risk for heavy smoking. *Mol Psychiatry* 13(4):368–373.

Bierut LJ, Rice JP, Goate A, Hinrichs AL, Saccone NL, Foroud T, Edenberg HJ, Cloninger CR, Begleiter H, Conneally PM and others. 2004. A genomic scan for habitual smoking in families of alcoholics: common and specific genetic factors in substance dependence. *Am J Med Genet A* 124A(1):19–27.

Bierut LJ, Madden PA, Breslau N, Johnson EO, Hatsukami D, Pomerleau OF, Swan GE, Rutter J, Bertelsen S, Fox L and others. 2007. Novel genes identified in a high-density genome wide association study for nicotine dependence. *Hum Mol Genet* 16(1):24–35.

Bierut LJ, Agrawal A, Bucholz KK, Doheny KF, Laurie C, Pugh E, Fisher S, Fox L, Howells W, Bertelsen S and others. 2010. A genome-wide association study of alcohol dependence. *Proc Natl Acad Sci USA* 107(11):5082–5087.

Bowden CL, Calabrese JR, McElroy SL, Gyulai L, Wassef A, Petty F, Pope HG, Jr., Chou JC, Keck PE, Jr., Rhodes LJ and others. 2000. A randomized, placebo-controlled 12-month trial of divalproex and lithium in treatment of outpatients with bipolar I disorder. Divalproex Maintenance Study Group. *Arch Gen Psychiatry* 57(5):481–489.

Butt CM, King NM, Stitzel JA, Collins AC. 2004. Interaction of the nicotinic cholinergic system with ethanol withdrawal. *J Pharmacol Exp Ther* 308(2):591–599.

Campbell-Scherer D. 2010. Multimorbidity: a challenge for evidence-based medicine. *Evid Based Med* 15(6):165–166.

Caporaso N, Gu F, Chatterjee N, Sheng-Chih J, Yu K, Yeager M, Chen C, Jacobs K, Wheeler W, Landi MT and others. 2009. Genome-wide and candidate gene association study of cigarette smoking behaviors. *PLoS One* 4(2):e4653.

Clark A, Lindgren S, Brooks SP, Watson WP, Little HJ. 2001. Chronic infusion of nicotine can increase operant self-administration of alcohol. *Neuropharmacology* 41(1):108–117.

de Groot V, Beckerman H, Lankhorst GJ, Bouter LM. 2003. How to measure comorbidity. A critical review of available methods. *J Clin Epidemiol* 56(3):221–229.

DeLong ER, DeLong DM, Clarke-Pearson DL. 1988. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* 44(3):837–845.

Dilsaver SC, Benazzi F, Oedegaard KJ, Fasmer OB, Akiskal HS. 2009a. Is a family history of bipolar disorder a risk factor for migraine among affectively ill patients? *Psychopathology* 42(2):119–123.

Dilsaver SC, Benazzi F, Oedegaard KJ, Fasmer OB, Akiskal KK, Akiskal HS. 2009b. Migraine headache in affectively ill latino adults of Mexican American origin is associated with bipolarity. *Prim Care Companion J Clin Psychiatry* 11(6):302–306.

Ehringer MA, Clegg HV, Collins AC, Corley RP, Crowley T, Hewitt JK, Hopfer CJ, Krauter K, Lessem J, Rhee SH and others. 2007. Association of the neuronal nicotinic receptor beta2 subunit gene (CHRNB2) with subjective responses to alcohol and nicotine. *Am J Med Genet B Neuropsychiatr Genet* 144B(5): 596–604.

Feinstein A. 1970. The pre-therapeutic classficaitoin of co-morbidity in chronic disease. *J Chronic Dis* 23(7):455–468.

Ferreira MA, O'Donovan MC, Meng YA, Jones IR, Ruderfer DM, Jones L, Fan J, Kirov G, Perlis RH, Green EK and others. 2008. Collaborative genome-wide association analysis supports a role for ANK3 and CACNA1C in bipolar disorder. *Nat Genet* 40(9):1056–1058.

Gijsen R, Hoeymans N, Schellevis FG, Ruwaard D, Satariano WA, van den Bos GA. 2001. Causes and consequences of comorbidity: a review. *J Clin Epidemiol* 54(7):661–674.

Grucza RA, Wang JC, Stitzel JA, Hinrichs AL, Saccone SF, Saccone NL, Bucholz KK, Cloninger CR, Neuman RJ, Budde JP and others. 2008. A risk allele for nicotine dependence in CHRNA5 is a protective allele for cocaine dependence. *Biol Psychiatry* 64(11):922–929.

Hoft NR, Corley RP, McQueen MB, Huizinga D, Menard S, Ehringer MA. 2009. SNPs in CHRNA6 and CHRNB3 are associated with alcohol consumption in a nationally representative sample. *Genes Brain Behav* 8(6):631–637.

Howie BN, Donnelly P, Marchini J. 2009. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet* 5(6):e1000529.

Howie B, Marchini J, Stephens M. 2011. Genotype imputation with thousands of genomes. *G3 (Bethesda)* 1(6):457–470.

John U, Meyer C, Rumpf HJ, Hapke U. 2003. Probabilities of alcohol high-risk drinking, abuse or dependence estimated on grounds of tobacco smoking and nicotine dependence. *Addiction* 98(6):805–814.

Larsson A, Engel JA. 2004. Neurochemical and behavioral studies on ethanol and nicotine interactions. *Neurosci Biobehav Rev* 27(8):713–720.

Lind PA, Macgregor S, Vink JM, Pergadia ML, Hansell NK, de Moor MH, Smit AB, Hottenga JJ, Richter MM, Heath AC and others. 2010. A genomewide association study of nicotine and alcohol dependence in Australian and Dutch populations. *Twin Res Hum Genet* 13(1):10–29.

Lu Q, Wei C, Ye C, Li M, Elston RC. 2012. A likelihood ratio based Mann-Whitney approach finds novel replicable joint gene action for type 2 diabetes. *Genet Epidemiol* 36:583–593.

Maj M. 2005. "Psychiatric comorbidity": an artefact of current diagnostic systems? *Br J Psychiatry* 186:182–184.

Marchini J, Donnelly P, Cardon LR. 2005. Genome-wide strategies for detecting multiple loci that influence complex diseases. *Nat Genet* 37(4):413–417.

Moskvina V, Craddock N, Holmans P, Nikolov I, Pahwa JS, Green E, Owen MJ, O'Donovan MC. 2009. Gene-wide analyses of genome-wide association data sets: evidence for multiple common risk alleles for schizophrenia and bipolar disorder and for overlap in genetic risk. *Mol Psychiatry* 14(3):252–260.

Neale MC, Kendler KS. 1995. Models of comorbidity for multifactorial disorders. *Am J Hum Genet* 57(4):935–953.

Oedegaard KJ, Greenwood TA, Johansson S, Jacobsen KK, Halmoy A, Fasmer OB, Akiskal HS, Haavik J, Kelsoe JR. 2010a. A genome-wide association study of bipolar disorder and comorbid migraine. *Genes Brain Behav* 9(7):673–680.

Oedegaard KJ, Greenwood TA, Lunde A, Fasmer OB, Akiskal HS, Kelsoe JR. 2010b. A genome-wide linkage study of bipolar disorder and co-morbid migraine: replication of migraine linkage on chromosome 4q24, and suggestion of an overlapping susceptibility region for both disorders on chromosome 20p11. *J Affect Disord* 122(1–2):14–26.

Rhee SH, Hewitt JK, Lessem JM, Stallings MC, Corley RP, Neale MC. 2004. The validity of the Neale and Kendler model-fitting approach in examining the etiology of comorbidity. *Behav Genet* 34(3):251–265.

Riala K, Hakko H, Isohanni M, Jarvelin MR, Rasanen P. 2004. Teenage smoking and substance use as predictors of severe alcohol problems in late adolescence and in young adulthood. *J Adolesc Health* 35(3):245–254.

Saccone SF, Hinrichs AL, Saccone NL, Chase GA, Konvicka K, Madden PA, Breslau N, Johnson EO, Hatsukami D, Pomerleau O and others. 2007. Cholinergic nicotinic receptor genes implicated in a nicotine dependence association study targeting 348 candidate genes with 3713 SNPs. *Hum Mol Genet* 16(1):36–49.

Schlaepfer IR, Hoft NR, Collins AC, Corley RP, Hewitt JK, Hopfer CJ, Lessem JM, McQueen MB, Rhee SH, Ehringer MA. 2008a. The CHRNA5/A3/B4 gene cluster variability as an important determinant of early alcohol and tobacco initiation in young adults. *Biol Psychiatry* 63(11):1039–1046.

Schlaepfer IR, Hoft NR, Ehringer MA. 2008b. The genetic components of alcohol and nicotine co-addiction: from genes to behavior. *Curr Drug Abuse Rev* 1(2):124–134.

Schuckit MA, Danko GP, Smith TL, Bierut LJ, Bucholz KK, Edenberg HJ, Hesselbrock V, Kramer J, Nurnberger JI, Jr., Trim R and others. 2008. The prognostic implications of DSM-IV abuse criteria in drinking adolescents. *Drug Alcohol Depend* 97(1–2):94–104.

Simonoff E. 2000. Extracting meaning from comorbidity: genetic analyses that make sense. *J Child Psychol Psychiatry* 41(5):667–674.

Sklar P, Smoller JW, Fan J, Ferreira MA, Perlis RH, Chambert K, Nimgaonkar VL, McQueen MB, Faraone SV, Kirby A and others. 2008. Whole-genome association study of bipolar disorder. *Mol Psychiatry* 13(6):558–569.

Spitz MR, Amos CI, Dong Q, Lin J, Wu X. 2008. The CHRNA5-A3 region on chromosome 15q24–25.1 is a risk factor both for nicotine dependence and for lung cancer. *J Natl Cancer Inst* 100(21):1552–1556.

Stevens VL, Bierut LJ, Talbot JT, Wang JC, Sun J, Hinrichs AL, Thun MJ, Goate A, Calle EE. 2008. Nicotinic receptor gene variants influence susceptibility to heavy smoking. *Cancer Epidemiol Biomarkers Prev* 17(12):3517–3525.

The International HapMap Consortium. 2003. The International HapMap Project. *Nature* 426(6968):789–796.

Thorgeirsson TE, Geller F, Sulem P, Rafnar T, Wiste A, Magnusson KP, Manolescu A, Thorleifsson G, Stefansson H, Ingason A and others. 2008. A variant associated with nicotine dependence, lung cancer and peripheral arterial disease. *Nature* 452(7187):638–642.

Treutlein J, Cichon S, Ridinger M, Wodarz N, Soyka M, Zill P, Maier W, Moessner R, Gaebel W, Dahmen N and others. 2009. Genome-wide association study of alcohol dependence. *Arch Gen Psychiatry* 66(7):773–784.

True WR, Xian H, Scherrer JF, Madden PA, Bucholz KK, Heath AC, Eisen SA, Lyons MJ, Goldberg J, Tsuang M. 1999. Common genetic vulnerability for nicotine and alcohol dependence in men. *Arch Gen Psychiatry* 56(7):655–661.

Valderas JM, Starfield B, Sibbald B, Salisbury C, Roland M. 2009. Defining comorbidity: implications for understanding health and health services. *Ann Fam Med* 7(4):357–363.

Wang JC, Grucza R, Cruchaga C, Hinrichs AL, Bertelsen S, Budde JP, Fox L, Goldstein E, Reyes O, Saccone N and others. 2009. Genetic variation in the CHRNA5 gene affects mRNA levels and is associated with risk for alcohol dependence. *Mol Psychiatry* 14(5):501–510.

Wang KS, Liu X, Zhang Q, Pan Y, Aragam N, Zeng M. 2011. A meta-analysis of two genome-wide association studies identifies 3 new loci for alcohol dependence. *J Psychiatr Res* 45(11):1419–1425.

Ware JJ, van den Bree MB, Munafo MR. 2011. Association of the CHRNA5-A3-B4 gene cluster with heaviness of smoking: a meta-analysis. *Nicotine Tob Res* 13(12):1167–1175.

Wellcome Trust Case Control Consortium. 2007. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* 447(7145):661–678.

Youngstrom EA, Arnold LE, Frazier TW. 2010. Bipolar and ADHD comorbidity: both artifact and outgrowth of shared mechanisms. *Clin Psychol (New York)* 17(4):350–359.

Zeiger JS, Haberstick BC, Schlaepfer I, Collins AC, Corley RP, Crowley TJ, Hewitt JK, Hopfer CJ, Lessem J, McQueen MB and others. 2008. The neuronal nicotinic receptor subunit genes (CHRNA6 and CHRNB3) are associated with subjective responses to tobacco. *Hum Mol Genet* 17(5):724–734.