Addictive Behaviors 37 (2012) 1138–1144

Contents lists available at SciVerse ScienceDirect

Addictive Behaviors

Improved methods to identify stable, highly heritable subtypes of opioid use and related behaviors

Jiangwen Sun [a], Jinbo Bi [a], Grace Chan [b], David Oslin [c], Lindsay Farrer [d], Joel Gelernter [e], Henry R. Kranzler [c,*]

[a] Department of Computer Science and Engineering, University of Connecticut, Storrs, CT USA
[b] Department of Psychiatry, University of Connecticut School of Medicine, Farmington, CT, USA
[c] Department of Psychiatry, University of Pennsylvania School of Medicine and the VISN 4 MIRECC, Philadelphia VAMC, Philadelphia, PA, USA
[d] Departments of Medicine (Genetics Program), Neurology, Genetics & Genomics, Epidemiology, and Biostatistics, Boston University Schools of Medicine and Public Health, Boston, MA, USA
[e] Departments of Psychiatry, Genetics, and Neurobiology, Yale University School of Medicine, New Haven, CT and VA CT Healthcare Center, West Haven, CT, USA

ARTICLE INFO

Keywords:
Opioid dependence
Subtypes
Phenotype
k-medoids clustering
Hierarchical clustering
Heritability

ABSTRACT

Although there is evidence that opioid dependence (OD) is heritable, efforts to identify genes contributing to risk for the disorder have been hampered by its complex etiology and variable clinical manifestations. Decomposition of a complex set of opioid users into homogeneous subgroups could enhance genetic analysis. We applied a series of data mining techniques, including multiple correspondence analysis, variable selection and cluster analysis, to 69 opioid-related measures from 5390 subjects aggregated from family-based and case–control genetic studies to identify homogeneous subtypes and estimate their heritability. Novel aspects of this work include our use of 1) heritability estimates of specific clinical features of OD to enhance the heritability of the subtypes and 2) a k-medoids clustering method in combination with hierarchical clustering to yield replicable clusters that are less sensitive to noise than previous methods. We identified five homogeneous groups, including two large groups comprised of 762 and 1353 heavy opioid users, with estimated heritability of 0.69 and 0.76, respectively. These methods represent a promising approach to the identification of highly heritable subtypes in complex, heterogeneous disorders.

Published by Elsevier Ltd.

1. Introduction

Opioid dependence (OD) is a serious, prevalent disorder and the number of people using opioids in the United States has been increasing since 1990 (Anthony, Warner, & Kessler, 1994; Regier et al., 1990; Substance Abuse and Mental Health Services Administration, 2007). In 2006, 0.8% of Americans aged 12 or older met criteria for a lifetime opioid use disorder and 0.4% were treated for an opioid-related problem in the past year (Substance Abuse and Mental Health Services Administration, 2007).

The etiology of OD is complex and multifactorial with a substantial heritable component (Gelernter et al., 2006; Kendler, Jacobson, Prescott, & Neale, 2003; Saxon, Oreskovich, & Brkanac, 2005). However, the heterogeneous phenotype defined by the DSM-IV diagnosis of OD (American Psychiatric Association, 1994) does not lend itself readily to gene finding (Gelernter et al., 2006). Thus, the identification of valid and homogeneous subgroups based on opioid use and related behaviors can refine the phenotype and enhance genetic analysis. This subtyping approach could facilitate the development of new treatments targeted to specific subgroups, enhancing personalized care (Chan, Gelernter, Oslin, Farrer, & Kranzler, 2011; Gelernter et al., 2006; Kranzler et al., 2008; Sabb et al., 2009). In the present study, we sought to identify opioid use subgroups that 1) differed significantly on clinical features and 2) demonstrated high heritability. Heritability is the ratio of additive genetic variance to the total phenotypic variance within a population (Bell, 1977). Although the heritability of a trait reflects the genetic contribution to that trait in the population, it cannot be directly applied to estimate the likelihood that an individual in that population will have the trait.

Empirical subtyping approaches rest on theories that emphasize the multifaceted nature of substance use and related behaviors (Basu, Ball, Feinn, Gelernter, & Kranzler, 2004; Moss, Chen, & Yi, 2007). Due to the complexity of these phenomena, empirical subtyping approaches have outperformed traditional methods of subtyping (Ball, Carroll, Babor, & Rounsaville, 1995; Epstein, Labouvie, McCrady, Jensen, & Hayaki, 2002). Although a variety of univariate empirical subtyping approaches have been used (Ball et al., 1995; Craig & Olson, 1992), multivariate cluster analysis has been the method of choice to subtype substance dependence (Chan et al., 2011; Gelernter et al., 2005, 2006; Kranzler et al., 2008). This approach has been used successfully to identify subtypes of cocaine dependence (CD) (Kranzler et al., 2008) and opioid dependence (OD) (Chan et al., 2011). These subtypes allowed the identification of

* Corresponding author at: Department of Psychiatry, University of Pennsylvania School of Medicine, Treatment Research Center, 3900 Chestnut St., Philadelphia, PA 19104, USA. Tel.: +1 215 222 3200x137; fax: +1 215 386 6770.
E-mail address: kranzler_h@mail.trc.upenn.edu (H.R. Kranzler).

0306-4603/$ – see front matter. Published by Elsevier Ltd.
doi:10.1016/j.addbeh.2012.05.010

promising candidate regions in genome-wide linkage scans for CD (Gelernter et al., 2005) and OD (Gelernter et al., 2006). The approach used a range of cocaine or opioid use behaviors to which k-means (Seber, 1984), a classic non-hierarchical clustering method, was applied in combination with a hierarchical clustering method (Hastie, Tibshirani, & Friedman, 2001) based on Ward's aggregation criterion. However, because the k-means method is an iterative procedure initialized with randomly chosen cluster centers, it is sensitive to outliers, with different initialization known to yield different clusters. Cross tagging multiple k-means mitigates this problem, but does not guarantee the creation of replicable subgroups.

The current study was designed to address the limitations of prior studies, including our own. This effort consisted in revising the statistical analytic approach to include a variable selection step that identifies the characteristic features of opioid use and related behaviors that are most likely to be heritable to guide the creation of clusters. The analytic approach also replaced the multiple k-means method with a number of random initiations by the k-medoids method with a deterministic initiation, thus ensuring the replicability of the resultant clusters.

2. Method

2.1. Subject recruitment

A total of 5390 subjects were aggregated from family-based and case–control genetic studies of DSM-IV OD and CD (American Psychiatric Association, 1994). Subjects were recruited at five sites: the University of Connecticut Health Center (n = 2224), Yale University School of Medicine (n = 2210), the University of Pennsylvania School of Medicine (n = 477), McLean Hospital (n = 258) and the Medical University of South Carolina (n = 221). The institutional review board at each site approved the protocol and informed consent forms. The National Institute on Drug Abuse provided a Certificate of Confidentiality. Subjects were paid for their participation.

The sample set included 3328 unrelated individuals and 2062 subjects from 864 small nuclear families, including all of the 4061 participants in the study by Chan et al. (2011). Of the families, 356 (41.2%) had ≥2 members with OD, including 229 families (26.6%) with ≥2 members having both OD and CD. Overlapping these were 701 (81.3%) families with ≥2 members with CD [thus 327 (37.9%) of the CD families had ≥2 members with CD but not OD]. Additionally, 54 (6.3%) families had only one affected member. Viewed differently, there were 864 probands, 1073 siblings, 95 parents, and 30 other family members from family-based studies, and 2685 subjects with OD and/or CD and 643 controls from case–control studies. Pedigree information was obtained for all small nuclear families that were recruited. Control subjects were screened to exclude those with a lifetime substance use disorder. Subjects with a clinical diagnosis of a psychotic disorder or gross cognitive impairment were excluded.

The majority of subjects (57.3%) were never married, 27.5% were widowed, separated, or divorced, and 15.2% were married at the time of the interview. The self-reported ethnic/racial distribution of the sample was 46.6% African-American (AA), 35.4% European-American (EA), 9.4% Hispanic, and 8.6% Native American, Pacific Islander or members of other minority groups. With respect to the level of education, 5.4% completed grade school only; 34.8% had some high school, but no diploma; 27.9% had completed high school; and 31.9% received education beyond high school.

2.2. Assessments

Subjects were assessed with the semi-structured assessment for drug dependence and alcoholism (SSADDA), a computer-assisted interview that yields lifetime DSM-IV diagnoses of substance use and most Axis I psychiatric disorders, as well as antisocial personality disorder (Pierucci-Lagha et al., 2005, 2007). It includes a specific section dedicated to the diagnosis of OD. The test–retest and inter-rater reliabilities of the SSADDA diagnosis of OD were excellent, with $\kappa = 0.94$ and 0.91, respectively (Pierucci-Lagha et al., 2005).

More than 43% of the subjects (2320) had a lifetime DSM-IV diagnosis of OD (1404 men and 916 women). The three most common diagnoses were CD (76.5%), nicotine dependence (62.4%) and alcohol dependence (46.2%). Major depressive episode (MDE) was the most common psychiatric disorder (15.6%), followed by posttraumatic stress disorder (PTSD) (14.6%), antisocial personality disorder (ASPD) (12.7%) and pathological gambling (8.8%). Generalized estimating equations (GEE) Wald Type 3 χ^2-tests with Bonferroni correction for multiple comparisons showed that men were significantly more likely than women to have a diagnosis of dependence on cocaine ($\chi^2_{(1)} = 39.3$, $p < 0.001$), alcohol ($\chi^2_{(1)} = 72.3$, $p < 0.001$), opioids ($\chi^2_{(1)} = 61.7$, $p < 0.001$) and other substances ($\chi^2_{(1)} = 69.1$, $p < 0.001$), antisocial personality disorder (ASPD) ($\chi^2_{(1)} = 121.9$, $p < 0.001$) and compulsive gambling ($\chi^2_{(1)} = 112.4$, $p < 0.001$). Women were more likely to receive a diagnosis of a major depressive episode ($\chi^2_{(1)} = 95.1$, $p < 0.001$), posttraumatic stress disorder (PTSD) ($\chi^2_{(1)} = 41.5$, $p < 0.001$), obsessive–compulsive disorder (OCD) ($\chi^2_{(1)} = 18.8$, $p < 0.001$), agoraphobia ($\chi^2_{(1)} = 60.2$, $p < 0.001$) and panic disorder ($\chi^2_{(1)} = 35.5$, $p < 0.001$).

2.3. Measures

The opioid drug section of the SSADDA contains 23 questions on (1) age of onset, frequency, and intensity of opioid use; (2) route of opioid administration; (3) occurrence of psychosocial and medical consequences of opioid use; (4) attempts to quit opioid use; and (5) opioid treatment history, resulting in 220 variables. A previous study by Chan et al. (2011) using a subset of the sample used in the present study identified key questions in this section for the purpose of subtyping opioid use and related behaviors. Those features were based on the apparent clinical utility of the features as discriminators of opioid-use-behavior subtypes (Chan et al., 2011).

Supplementary Tables 1 and 2 provide a complete list of the 69 key variables from the OD section of the SSADDA, which were used to generate clusters. Demographics and other substance use and psychiatric variables and disorders obtained from the SSADDA interview, together with heritability estimates, were used to evaluate the validity of the clusters.

The majority (i.e., 55) of the 69 key variables were categorical, with four possible response categories: “yes”, “no”, “obligate no”, and “missing.” For the 5390 participants, complete data were available for 98.93% of the entries for the 69 key variables. Fifteen of the key variables asked about the signs and symptoms that a subject experienced when he/she stopped, reduced, or went without opioids. Using “yes-or-no” questions, respondents were asked whether they had *ever* experienced the 15 withdrawal symptoms (“ever occur” symptoms) and whether two or more symptoms occurred *together* (“occur together” symptoms). If participants never tried to stop or reduce their opioid use, these variables were scored as “obligate no.” Our previous subtyping efforts (Chan et al., 2011; Gelernter et al., 2005, 2006; Kranzler et al., 2008) used the set of “ever occur” withdrawal variables. In the present study, based on the analysis described below, we used the “occur together” withdrawal variables.

2.4. Data analysis

Our analysis comprised three steps (see Supplementary Fig. 1): data reduction, cluster analysis, and heritability estimation. First, we used variable selection (Guyon & Elisseeff, 2003) and multiple correspondence analysis (MCA) (Abdi & Valentin, 2007; LeRoux & Rouanet, 2009; Murtagh, 2007) to reduce the large number of variables. In the variable selection step, we focused on the selection of withdrawal signs and symptoms, which was the largest subset of variables in

the analysis. The MCA data reduction approach is similar to principal components analysis but it compacts *categorical* (rather than continuous) data to a lower-dimensional space (Greenacre & Hastie, 1987). The retained principal dimensions are those that explain substantial variance in the data. The output of MCA comprised the coordinates of the retained dimensions for each of the 5390 subjects. MCA was first used to find the principal dimensions for the 15 "occur together" symptoms and the 15 "ever occur" symptoms, respectively. The variability and heritability of these two sets of principal dimensions were compared to select between the two sets of variables. MCA was then applied to all of the 69 selected variables to reduce the data dimension. The number of dimensions retained was guided by the Benzécri adjusted cumulative percentage, showing the percentage of variance explained by the retained dimensions (Benzecri, 1992).

Second, we used cluster analysis, which groups similar subjects together based on their clinical features, to create clusters of subjects. In the present study, we combined the k-medoids clustering method (Kaufman & Rousseeuw, 1990; Theodoridis & Koutroumbas, 1999; van der Laan, Pollard, & Bryan, 2003) consecutively with agglomerative hierarchical clustering (Calinski & Harabasz, 1974; Day & Edelsbrunner, 1984; Milligan, 1979; Tan, Steinbach, & Kumar, 2006). The k-medoids method first partitioned the subjects into 100 intermediate clusters. Then hierarchical clustering was used to merge the intermediate clusters to form a hierarchy of clusters based on Ward's aggregation criterion, yielding a dendrogram and statistics such as cubic clustering criterion (CCC), R^2, pseudo F and pseudo t^2, which guided the determination of the final number of clusters. To produce more reliable clusters, the clustering approach used here differs in a number of ways from the k-means approach of Chan et al. (2011). Specifically, rather than using the average of subjects in a cluster as the cluster centroid, the k-medoids method groups data by finding the most representative subjects to serve as cluster centroids. Thus, the subject whose measures were the closest (having the least sum of distances) to the measures of all other subjects was selected as the first representative (Kaufman & Rousseeuw, 1990). Subsequently, subjects were selected to increase the within-cluster similarity until k representative subjects were chosen as the initial cluster centroids. Once the initialization was completed, k-medoids iteratively exchanged selected representatives with unselected ones to improve the within-cluster similarity.

We used SAS 9.2 (SAS Institute Inc., 2008) to conduct the data reduction and cluster analysis, and the partitioning around medoids (PAM) package in the R language (Calinski & Harabasz, 1974; Kaufman & Rousseeuw, 1990) for the k-medoids method. After determining the final number of clusters, we characterized the resultant clusters using 33 variables reflecting demographics, opioid use behaviors, and related non-opioid use behaviors. The characteristics of each cluster were used to label the clusters. GEE Wald Type 3 χ^2-tests were used to determine whether the clusters differed significantly on these variables. We used Bonferroni correction ($p<0.05/33=0.0015$) to avoid inflating the Type I error rate.

To estimate the heritability of each of the clusters, logistic regression was first used to construct a classifier to separate subjects in each of the different clusters. The resultant classifier, as a function of the 69 measures of opioid use and related behaviors, calculated the likelihood that each subject belonged to a specific cluster. The log likelihood of 4964 subjects from EA and AA populations with 1805 of them from multi-member families was submitted to Sequential Oligogenic Linkage Analysis Routines (SOLAR) (Almasy & Blangero, 1998) software together with pedigrees to estimate the heritability of the cluster-derived trait. Including singleton cases together with multi-member families in the heritability estimation helped to correct the bias in the family-based sample due to the ascertainment method and is the preferred approach (Almasy & Blangero, 1998). Sex, age, and race were used as covariates in the heritability estimate.

3. Results

Because few participants endorsed each of the individual "ever occur" and "occur together" withdrawal symptoms, we reduced these sparse variables into fewer principal dimensions using MCA. Table 1 shows the first three MCA dimensions that together explain more than 80% of the variance for both the "ever occur" and "occur together" symptoms. The three dimensions were evaluated for all subjects.

Only the first MCA dimension for each of the two variable sets showed substantial heritability, with this dimension of the "occur together" symptoms being more informative than the "ever occur" symptoms. Further, because the first MCA dimension for the "ever occur" symptoms did not vary among individuals with lifetime OD it did not help to differentiate the OD subtypes. On this basis, we chose to use the 15 "occur together" symptoms in the cluster analysis in conjunction with 54 previously selected variables (listed in Supplementary Tables 1 and 2). The MCA reduced the total 69 categorical variables to 10 continuous dimensions, explaining over 99% of the variance.

The k-medoids cluster analysis partitioned the 5390 subjects into 100 mutually exclusive clusters based on the 10 MCA output dimensions. The hierarchical clustering method aggregated the 100 clusters into a hierarchy from 1 to 100, producing a dendrogram with pseudo F and t^2 statistics. These statistics suggested that between 3 and 8 clusters were optimal for this sample. In the final step, we visually inspected the features of the clusters to identify the clinically different characteristics of the clusters. Five mutually exclusive clusters were finally identified. As shown in Table 2, these clusters (subtypes) differed significantly on age, sex, race, education and marital status. Specifically, Groups 2–5 included significantly more men than women and these groups were less educated than Group 1. Group 5 had the lowest level of education. Groups 3 to 5 included a significantly higher proportion of EAs than the other two groups and Group 5 had the fewest married participants.

3.1. Opioid use and related behaviors by subgroup

We evaluated the validity of our five-cluster solution by comparing the clusters on the lifetime prevalence of substance use and psychiatric disorders (Table 3) and on opioid-related features (Table 4), as summarized below.

3.1.1. Non-opioid users (Group 1)

Of these 2,756 individuals (51.1% of the sample), 52.8% were women, and 62.6% African-American. Less than 23% of the subjects

Table 1
Comparison of the MCA dimensions for "ever occur" and "occur together" opioid withdrawal symptoms.

MCA dimension	"Ever occur" symptoms			"Occur together" symptoms		
	First[a]	Second[b]	Third[c]	First[a]	Second[b]	Third[c]
Percent age of total variance explained	33.3	33.0	20.5	30.4	26.7	23.8
Variance in subjects without lifetime OD	0.376	1.45	0.416	0.035	0.913	0.053
Variance in subjects with OD	0	0.382	0.748	0.354	0.502	0.818
Heritability	0.60	0	0.03	0.75	0.002	0.06

MCA, multiple correspondence analysis; OD, opioid dependence.
[a] The first MCA Dimension is the factor that explains the largest percentage of sample variance in the 69 OD variables.
[b] The second MCA dimension is the factor that explains the largest percentage of variance remaining after that explained by the first dimension is removed.
[c] The third MCA dimension is the factor that explains the largest percentage of the variance remaining after that explained by the first and second dimensions is removed.

Table 2
Demographic characteristics by group.

Characteristic[b]	Group 1 2756(51.1)	Group 2 391(7.3)	Group 3 128(2.4)	Group 4 762(14.1)	Group 5 1353(25.1)	Test statistic[a]
Age [mean(SD)]	41.1(9.7)	40.1(9.8)	49.1(5.3)	40.1(9.0)	37.9(9.6)	$\chi^2_{(4)} = 388.92$
Sex [*N*(%)]						$\chi^2_{(4)} = 118.63$
Women	1454(52.8)	128(32.7)	53(41.4)	326(42.8)	504(37.3)	
Men	1302(47.2)	263(67.3)	75(58.6)	436(57.2)	849(62.8)	
Race [*N*(%)]						$\chi^2_{(12)} = 620.50$
AA	1724(62.6)	179(45.8)	37(28.9)	245(32.2)	327(24.2)	
EA	669(24.3)	147(37.6)	69(53.9)	340(44.6)	681(50.3)	
Hispanic	162(5.9)	31(7.9)	11(8.6)	106(13.9)	198(14.6)	
Other	200(7.3)	34(8.7)	11(8.6)	71(9.3)	147(10.9)	
Education [*N*(%)]						$\chi^2_{(12)} = 189.68$
No HS	106(3.9)	19(4.9)	8(6.3)	47(6.2)	113(8.4)	
Some HS	811(29.4)	152(38.9)	35(27.3)	305(40.0)	570(42.1)	
HS graduate	758(27.5)	103(26.3)	43(33.6)	221(29.0)	376(27.8)	
Beyond HS	1080(39.2)	117(29.9)	42(32.8)	188(24.7)	293(21.7)	
Marital status [*N*(%)]						$\chi^2_{(8)} = 102.83$
Never married	1528(55.4)	233(59.6)	45(35.2)	431(56.6)	850(62.8)	
Married	522(18.9)	49(12.5)	19(14.8)	89(11.7)	140(10.4)	
Div/Sep/Wid	706(25.6)	109(27.9)	64(50.0)	242(31.8)	363(26.8)	

AA: African American, EA: European American; HS: high school; Div/Sep/Wid: divorced, separated, or widowed.
[a] All demographic variables differed significantly by group at $p<0.0015$ (i.e., Bonferroni correction: $p<0.05/33$).
[b] GEE Wald χ^2 tests for age and sex; independent χ^2 tests for race, education and marital status because the GEE.

had ever used an opioid, with a mean number of lifetime opioid uses of only 3.8 (SD = 3.0). Although no one in this group had a diagnosis of OD, many of them met criteria for other lifetime drug dependence disorders, notably CD (70.4%), nicotine dependence (49.4%), and alcohol dependence (39.9%). Nevertheless, this group had the lowest prevalence of all other substance dependence and psychiatric disorders except ASPD and social phobia.

3.1.2. Low-to-moderate opioid users (Group 2)

This group consisted of 391 individuals, 7.3% of the sample. Lifetime OD was significantly lower (25.1%) in this group than in Groups 3–5. Other than Group 1, Group 2 also had the lowest percentage of daily or almost daily opioid use (49.1%) and injection opioid use (20%), and the lowest percentage of subjects with negative effects due to opioid use and who ever received opioid treatment (Table 4). Daily expenditures for opioids were significantly lower in this group than in Groups 4 and 5 (GEE Wald $\chi^2_{(2)} = 197.2$, $p<0.001$, Group 3: mean (SD) = $77.6 (87.7), Group 4: mean (SD) = $104.3 (127.5), Group 5: mean (SD) = $148.7 (145.4)). Subjects in this group were also the least likely to have nicotine dependence, sedative dependence, PTSD, OCD and panic disorder. This group, however, had the highest prevalence of CD (87.2%), and alcohol dependence (57.3%).

3.1.3. Late-onset heavy opioid users (Group 3)

With only 128 individuals (2.4% of the total sample), this was the smallest group. Subjects in this group had significantly later onset of first [mean age (SD) = 33.7 (10.3)] and heaviest opioid use [mean age (SD) = 43 (5.1)] than those in the other groups. Nearly all subjects in this group received a diagnosis of OD, but significantly fewer subjects injected opioids than in Groups 4 and 5. This group had the highest rate of panic disorder but the lowest prevalence of CD, ASPD, social phobia and agoraphobia.

3.1.4. Heavy opioid users (Group 4)

This group comprises 762 individuals (14.1% of the sample), nearly 97% of whom had lifetime OD. The ages of first opioid use [mean (SD) = 23.3 (7.4)] and onset of heaviest use [mean (SD) = 29 (8.2)] were both intermediate, earlier than Group 3 but later than Group 5. The proportion experiencing negative effects due to opioid use and that received treatment for opioid abuse was similar to that in Group 3 but significantly lower than in Group 5. The prevalence of other substance dependence and psychiatric disorders was also significantly lower in this group than in Group 5.

3.1.5. Early-onset, highly comorbid, heavy opioid users (Group 5)

The 1353 subjects in this group (25.1% of the total sample) were the heaviest substance users, and were significantly affected by their opioid use: 75.1% reported arrests or trouble with police due to opioid use, significantly higher than all other groups. Subjects reported both the earliest onset of opioid use [mean age (SD) = 18.9 (4.2)] and the heaviest use [mean (SD) = 25.4 (7.0)]. They also had the highest prevalence of other substance dependence disorders (except cocaine and alcohol dependence) and psychiatric disorders (except panic disorder).

3.2. Heritability

The two heavy opioid user clusters, Groups 4 and 5, were the largest among the four opioid user groups and had the highest estimated heritability: 0.69 (SE = 0.06) and 0.76 (SE = 0.05), respectively (p's$<10^{-30}$). The heritability of the other two user groups, Groups 2 and 3, was also relatively high: 0.49 (SE = 0.07) and 0.53 (SE = 0.06), respectively (p's$<10^{-12}$). The non-opioid user Group 1 showed a heritability of 0.62 (SE = 0.06, $p<10^{-18}$). Race was a highly significant covariate in all groups (p's ranging from $<10^{-43}$ to 10^{-84}), sex was a significant covariate in all groups except Group 2 (p's ranging from $<10^{-6}$ to 10^{-19}), and age was a significant covariate in Groups 1, 3 and 5 (p's ranging from $<10^{-6}$ to 10^{-7}). Our sensitivity analysis on heritability estimation using different combinations of covariates (Supplementary Table 3) showed that the inclusion of race as a covariate reduced the estimated heritability by about 0.1 for all groups.

3.3. Comparison of subtypes derived using different cluster methods

The study by Chan et al. used a three-step LMW clustering procedure to identify OD subgroups in a sample of 4061 subjects, all of which were included in the present study (Chan et al., 2011). To differentiate whether the results reported here differed from those of Chan et al. due to the larger sample in the current study or the different analytic approach employed, we applied the LMW approach of Chan et al. to the full sample of 5390 subjects. This also resulted in 5 clusters. Table 5 cross-tabulates Groups A to E from the LMW

Table 3
Lifetime prevalence of substance use and psychiatric disorders by group [N (%)].

Disorder[a]	Group 1 2756(51.1)	Group 2 391(7.3)	Group 3 128(2.4)	Group 4 762(14.1)	Group 5 1353(25.1)	$\chi^2_{(4)}$ Test statistic[b]
Substance use disorders						
Cocaine dependence	1940(70.4)	341(87.2)	94(73.4)	592(77.7)	1154(85.3)	142.98
Nicotine dependence	1362(49.4)	267(68.3)	93(72.7)	548(71.9)	1096(81.0)	409.43
Alcohol dependence	1100(39.9)	224(57.3)	64(50.0)	375(49.2)	728(53.8)	103.70
Opioid dependence	0(0)	98(25.1)	126(98.4)	745(97.8)	1351(99.9)	444.78
Sedative dependence	35(1.3)	14(3.6)	10(7.8)	83(10.9)	242(17.9)	256.02
Stimulant dependence	77(2.8)	31(7.9)	10(7.8)	46(6.0)	113(8.4)	61.10
Other substance dependence[c]	97(3.5)	36(9.2)	12(9.4)	197(25.9)	537(39.7)	635.56
Psychiatric disorders						
ASPD[c]	252(9.1)	60(15.4)	9(7.0)	112(14.7)	249(18.4)	78.58
MDE[c]	368(13.4)	67(17.1)	23(18.0)	126(16.5)	257(19.0)	23.51
PTSD[c]	347(12.6)	51(13.0)	25(19.5)	117(15.4)	246(18.2)	25.78
OCD[c]	39(1.4)	2(1.6)	12(3.1)	21(2.8)	48(3.6)	20.13
Social phobia	68(2.5)	15(3.8)	3(2.3)	30(3.9)	87(6.4)	37.51
Agoraphobia	115(4.2)	21(5.4)	6(4.7)	39(5.1)	119(8.8)	38.17
Panic disorder	81(3.0)	21(5.4)	18(14.1)	50(6.6)	169(12.5)	136.22
Compulsive gambling	205(7.4)	37(9.5)	13(10.2)	69(9.1)	152(11.2)	17.70

ASPD: antisocial personality disorder; MDE: major depressive episode; PTSD: posttraumatic stress disorder; OCD: obsessive–compulsive disorder.
[a] All disorders differed significantly by group at $p<0.0015$ (i.e., Bonferroni correction: $p<0.05/33$).
[b] GEE Wald χ^2 tests for all disorders.
[c] Other substance dependence includes dependence on phencyclidine, hallucinogens, inhalants, solvents, or a combination of opioids and cocaine (i.e., "speedballs").

analysis with Groups 1 to 5 derived using our analytic method. The moderate opioid user Group B and the heavy late-onset user Group C resulting from the LMW method (estimated heritability = 0.70 and 0.60) comprised only 449 and 88 individuals, respectively, and the estimated heritability of all other groups was below 0.50. The groups that had an early-onset of OD (i.e., highly comorbid and heavy opioid users) largely overlapped in Group E and Group 5. However, the substantial disagreement between Group E in Chan et al. (2011) and Group 5 in the present analysis resulted in a large difference in heritability estimates for these groups. Groups 3 and 4 in our solution were split into Groups A to E using the LMW approach. A substantial minority (12.1%) of subjects from Group A (low-level or non-opioid users) with a lifetime OD diagnosis clustered (appropriately) in the opioid user groups in our analysis. Specifically, 332 (10.6%) individuals that were in Group A by the LMW method and in Groups 4 and 5 of our solution received an OD diagnosis and used opioids daily or almost daily, with the majority (85%) having stayed high for a whole day or more. Thus, it was inappropriate to identify these heavy users as low-level or non-opioid users. Further, inclusion of these heavy users in the low-level or non-opioid user group increased its phenotypic variance, thereby reducing its heritability.

4. Discussion

This study showed that carefully selected analytic methods enhance the validity and potential utility of empirically derived subtypes based on opioid-use behaviors. The subtypes are the result of multivariate analyses, so that the choice of one or a few parameters on which to compare them cannot adequately capture the differences among subtypes. Although these methods cannot readily be applied in a clinical setting, the findings presented here provide insight into subtypes that appear to have clinical significance. For example, as shown in Table 3, the subtypes differed significantly on a variety of co-occurring psychiatric disorders. This has important diagnostic and potential therapeutic implications that warrant further research.

A novel element of this study is that the variable selection used to generate clusters was guided by the heritability estimate for major features of opioid use related behaviors. Inclusion of the opioid withdrawal symptoms with a higher estimated heritability increased the heritability estimates of opioid use subtypes over those obtained previously by us using a similar approach. Extensions of this method can be used to examine other opioid-related measures or other disorders to yield a comprehensive set of informative and essential phenotypic features.

In addition to improving the analysis at the variable selection step, our approach differs from our previous studies (Chan et al., 2011; Gelernter et al., 2006) by replacing k-means cluster analysis, which uses several randomly chosen starting points, with a k-medoids method. Although repeating the k-means analyses with several starting points improves the stability of the resultant clusters, they may not be replicable at different runs of the clustering process. By using k-means analysis to create 50 clusters at each run and repeating it 10 times, 50^{10} cells have to be cross tagged to find stable clusters, requiring extensive computation. These 50^{10} cells may differ with different runs due to the randomness of starting points, leading to different cluster solutions. An information-theoretic criterion (Kaufman & Rousseeuw, 1990), such as the one used here, can select the initial points for the k-medoids analysis. Thus, the clusters derived using this approach do not vary when the analysis is run multiple times.

Consistent with our prior results (Chan et al., 2011; Gelernter et al., 2006), we identified five distinct subtypes in this sample of subjects participating in genetic studies of CD and OD. When we compared our results with those obtained in a previous analysis of a subset of these data (Chan et al., 2011), we found two groups that were larger and had higher heritability estimates than were obtained previously. Specifically, Groups 4 and 5 consisted of a total 2115 subjects, or 39% of the total sample, compared to only 984 subjects in two clusters with a high heritability estimate (24% of the total sample) in our previous study (Chan et al., 2011). This improves the potential utility of our approach for gene finding, by increasing the statistical power of studies that employ these subtypes. The groups in our solution were also phenotypically more distinct. For instance, our non-opioid-user group (Group 1) contained no subjects with a lifetime diagnosis of OD, compared with 20% of the lowest opioid-use group in our prior study (Chan et al., 2011). The late-onset group in our cluster solution had a significantly older age at first (33.7 years) and heaviest (43.0 years) opioid use than the late-onset group (first use at 26.6 years and heaviest use at 34.0 years) in Chan et al. (2011).

Because the more valid subtype analysis in the current study may have resulted from either a larger sample or a better subtyping method, we compared the two approaches by applying the LMW method of Chan et al. (2011) to the larger sample available for the present analysis. To do so, we used the same programs as in the prior study

Table 4
Lifetime opioid use characteristics, opioid-related effects, and opioid treatment history for Groups 1–5.[a]

Clinical features[b]	Group 2 391(7.3)	Group 3 128(2.4)	Group 4 762(14.1)	Group 5 1353(25.1)	$\chi^2_{(3)}$ Test statistic[c]
Opioid use characteristics					
Mean age of first opioid use in yr (SD)	23.3(7.9)	33.7(10.3)	23.28(7.4)	18.88(4.2)	536.37
Mean age of onset of heaviest opioid use in yr (SD)	27.9(9.1)	43.0(5.1)	28.95(8.2)	25.36(7.0)	1244.90
Used opioids daily or almost daily	192(49.1)	126(98.4)	737(96.7)	1346(99.5)	386.17
Injected opioids intravenously	78(20.0)	63(49.2)	433(56.8)	1045(77.2)	352.12
Opioid-related effects					
Stayed high from opioids for a whole day or more	193(49.4)	113(88.3)	638(83.7)	1253(92.6)	307.01
Strong desire for opioids made it hard to think of anything else	39(10.0)	109(85.2)	585(76.8)	1277(94.4)	594.53
Opioid use interfered with work, school, or home life	34(8.7)	90(70.3)	487(63.9)	1243(91.9)	581.31
Family members, friends, doctor, clergy, boss, or people at work or school objected to opioid use	66(16.9)	85(66.4)	572(75.1)	1288(95.2)	570.08
Been arrested or had trouble with the police because of opioid use	26(6.7)	60(46.9)	358(47.0)	1016(75.1)	395.58
Gave up or greatly reduced important activities due to opioid use	51(13.0)	96(75.0)	566(74.3)	1284(94.9)	628.27
Opioid treatment history					
Ever treated for an opioid-related problem	52(13.3)	110(86.0)	610(80.1)	1303(96.3)	655.70
Ever attended self-help group for opioid use	36(9.2)	73(57.0)	446(58.5)	1085(80.2)	406.36

[a] Values are N (%) of individuals endorsing the feature, unless otherwise specified. All subjects in Group 1 reported using opioids fewer than 11 times and thus skipped out of the rest of the opioid drugs section. This group was excluded from this table.
[b] All behaviors differed significantly by group at $p<0.0015$ (i.e., using Bonferroni correction: $p<0.05/33$).
[c] GEE Wald χ^2 tests for all behaviors.

(SAS and SOLAR, though the k-medoids analysis was run using the R package). We found that the current, modified approach produced not only larger clusters of higher heritability, but also more homogeneous clusters than our previous effort (Chan et al., 2011). Because the variable selection step that resulted in heritable OD measures led to more highly heritable subtypes, a thorough examination of the phenotypic measures used in subtyping methods may be necessary to optimize the procedure.

This study has a number of limitations. Due to lack of information about the childhood household of study participants, the estimated heritability may be inflated by shared environment among siblings. In the present study, substance dependent individuals were oversampled for genetic studies. Thus, the heritability of OD as estimated here is likely not to be representative of that in the general population. Additional sources of information concerning the identified subtypes, such as follow-up studies or molecular genetic correlation, are needed to validate these findings. The heritability estimates shown here are consistent with estimates for other substance dependence disorders. For example, it is estimated that alcohol dependence is 50–60% heritable and dependence on illicit drugs is 45–79% heritable (Dick & Agrawal, 2008). The heritability estimates for the subtypes in the present study are in this range, with the two most heritable subtypes at the high end of the range.

The high prevalence of CD in the study sample also limits our ability to generalize the findings to other OD samples without such comorbidity. Because the SSADDA does not provide details on the specific kinds of opioids that subjects used, there may be other subtypes of OD that are not captured in this study. Independent replication of our findings in a different sample is needed, as are studies using this approach to categorize other substance use and psychiatric features to yield homogeneous subtypes of other disorders. Such subtypes may have utility for gene finding and for clinical characterization and treatment selection. Despite the expectation that the identification of highly heritable subtypes of opioid use and related behaviors will enhance gene-finding efforts, this assumption must be tested empirically.

Role of funding sources
The funding sources did not directly influence the design or conduct of the study.

Contributors
Henry Kranzler, Joel Gelernter, Grace Chan, and Jinbo Bi designed the study. Henry Kranzler, Joel Gelernter, and David Oslin recruited the participants and oversaw their evaluation. Lindsay Farrer oversaw the data management and scoring of the phenotypic data. Henry Kranzler, Jiangwen Sun, and Jinbo Bi reviewed the literature and applied the findings to the current study. Jiangwen Sun, Jinbo Bi, and Grace Chan conducted the statistical analyses. Jiangwen Sun and Jinbo Bi wrote the first draft of the manuscript and all authors contributed to and approved the final manuscript.

Conflict of interest
Mr. Sun and Drs. Bi, Chan, Gelernter, and Oslin have no disclosures. Dr. Farrer received a research grant from Eisai Pharmaceuticals and consultant fees from Novartis Pharmaceuticals. Dr. Kranzler has been a paid consultant for Alkermes, Lilly, Lundbeck, Pfizer, and Roche. He also reports associations with Eli Lilly, Janssen, Schering Plough, Lundbeck, Alkermes, GlaxoSmithKline, Abbott, and Johnson & Johnson, as these

Table 5
Comparison of the subtyping approach in Chan et al. (2011) and our approach, as depicted in Supplementary Fig. 1. [N(%): number of overlapping subjects].

		Proposed approach				
		Group 1 (N = 2,756) $h^2 = 0.62$	Group 2 (N = 391) $h^2 = 0.49$	Group 3 (N = 128) $h^2 = 0.54$	Group 4 (N = 762) $h^2 = 0.69$	Group 5 (N = 1,353) $h^2 = 0.76$
Approach in Chan et al. (2011)	Group A (N = 3134) $h^2 = 0.29$	2756(87.94)	42(1.34)	4(0.13)	246(7.85)	86(2.74)
	Group B (N = 449) $h^2 = 0.70$		339(75.50)	4(0.89)	85(18.93)	21(4.68)
	Group C (N = 88) $h^2 = 0.60$			87(97.86)	1(1.14)	
	Group D (N = 434) $h^2 = 0.45$		10(2.30)	29(6.68)	334(76.96)	61(14.06)
	Group E (N = 1,285) $h^2 = 0.49$			4(0.31)	96(7.47)	1185(92.22)

companies provide support to the ACNP Alcohol Clinical Trials Initiative (ACTIVE) and he receives support from ACTIVE.

Acknowledgment

This work was supported by NIH grants DA12849, DA12690, DA22288, DA15105, DA005186, AA03510, AA11330, AA13736, and GM08607 and the VA CT and Philadelphia VA Mental Illness Research, Education, and Clinical Centers (MIRECCs).

Appendix A. Supplementary data

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.addbeh.2012.05.010.

References

Abdi, H., & Valentin, D. (2007). Multiple correspondence analysis. In N. Salkind (Ed.), *Encyclopedia of measurement and statistics* (pp. 651–657). Thousand Oaks: Sage.

Almasy, L., & Blangero, J. (1998). Multipoint quantitative-trait linkage analysis in general pedigrees. *The American Journal of Human Genetics*, *62*, 1198–1211.

American Psychiatric Association (1994). *Diagnosis and statistical manual of mental disorders (fourth edition)*. Washington DC: American Psychiatric Press, Inc.

Anthony, J., Warner, L., & Kessler, R. (1994). Comparative epidemiology of dependence on tobacco, alcohol, controlled substances, and inhalants: Basic findings from the National Comorbidity Survey. *Experimental and Clinical Psychopharmacology*, *2*(3), 244–268.

Ball, S. A., Carroll, K. M., Babor, T. F., & Rounsaville, B. J. (1995). Subtypes of cocaine abusers: Support for a type A-type B distinction. *Journal of Consulting and Clinical Psychology*, *63*(1), 115–124.

Basu, D., Ball, S. A., Feinn, R., Gelernter, J., & Kranzler, H. R. (2004). Typologies of drug dependence: Comparative validity of a multivariate and four univariate models. *Drug and Alcohol Dependence*, *73*(3), 289–300.

Bell, A. E. (1977). Heritability in retrospect. *Journal of Heredity*, *68*(5), 297–300.

Benzecri, J. P. (1992). *Correspondence analysis handbook*. New York: Marcel Dekker.

Calinski, T., & Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics — Simulation and Computation*, *3*, 1–27.

Chan, G., Gelernter, J., Oslin, D., Farrer, L., & Kranzler, H. R. (2011). Empirically derived subtypes of opioid use and related behaviors. *Addiction*, *106*(6), 1146–1154.

Craig, R. J., & Olson, R. (1992). MMPI subtypes for cocaine abusers. *The American Journal of Drug and Alcohol Abuse*, *18*(2), 197–205.

Day, W. H., & Edelsbrunner, H. (1984). Efficient algorithms for agglomerative hierarchical clustering methods. *Journal of Classification*, *1*, 7–24.

Dick, D. M., & Agrawal, A. (2008). The genetics of alcohol and other drug dependence. *Alcohol Research & Health*, *31*(2), 111–118.

Epstein, E. E., Labouvie, E., McCrady, B. S., Jensen, N. K., & Hayaki, J. (2002). A multi-site study of alcohol subtypes: Classification and overlap of unidimensional and multidimensional typologies. *Addiction*, *97*(8), 1041–1053.

Gelernter, J., Panhuysen, C., Weiss, R., Brady, K., Hesselbrock, V., Rounsaville, B., et al. (2005). Genomewide linkage scan for cocaine dependence and related traits: Significant linkages for a cocaine-related trait and cocaine-induced paranoia. *American Journal of Medical Genetics*, *136B*(1), 45–52.

Gelernter, J., Panhuysen, C., Wilcox, M., Hesselbrock, V., Rounsaville, B., Poling, J., et al. (2006). Genomewide linkage scan for opioid dependence and related traits. *The American Journal of Human Genetics*, *78*(5), 759–769.

Greenacre, M., & Hastie, T. (1987). The geometric interpretation of correspondence analysis. *Journal of the American Statistical Association*, *82*, 437–447.

Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research*, *3*, 1157–1182.

Hastie, T., Tibshirani, R., & Friedman, J. (2001). *The elements of statistical learning: Data mining, inference and prediction*. New York: Springer.

Kaufman, L., & Rousseeuw, P. J. (1990). *Finding groups in data: An introduction to cluster analysis*. New York: Wiley.

Kendler, K. S., Jacobson, K. C., Prescott, C. A., & Neale, M. C. (2003). Specificity of genetic and environmental risk factors for use and abuse/dependence of cannabis, cocaine, hallucinogens, sedatives, stimulants, and opiates in male twins. *The American Journal of Psychiatry*, *160*(4), 687–695.

Kranzler, H. R., Wilcox, M., Weiss, R. D., Brady, K., Hesselbrock, V., Rounsaville, B., et al. (2008). The validity of cocaine dependence subtypes. *Addictive Behaviors*, *33*(1), 41–53.

LeRoux, B., & Rouanet, H. (2009). *Multiple correspondence analysis*. Los Angeles: Sage.

Moss, H. B., Chen, C. M., & Yi, H. Y. (2007). Subtypes of alcohol dependence in a nationally representative sample. *Drug and Alcohol Dependence*, *91*(2–3), 149–158.

Milligan, G. W. (1979). Ultrametric hierarchical clustering algorithms. *Psychometrika*, *44*, 343–346.

Murtagh, F. (2007). Multiple correspondence analysis and related methods. *Psychometrika*, *72*, 275–277.

Pierucci-Lagha, A., Gelernter, J., Chan, G., Arias, A., Cubells, J. F., Farrer, L., et al. (2007). Reliability of DSM-IV diagnostic criteria using the semi-structured assessment for drug dependence and alcoholism (SSADDA). *Drug and Alcohol Dependence*, *91*(1), 85–90.

Pierucci-Lagha, A., Gelernter, J., Feinn, R., Cubells, J. F., Pearson, D., Pollastri, A., et al. (2005). Diagnostic reliability of the semi-structured assessment for drug dependence and alcoholism (SSADDA). *Drug and Alcohol Dependence*, *80*(3), 303–312.

Regier, D. A., Farmer, M. E., Rae, D. S., Locke, B. Z., Keith, S. J., Judd, L. L., et al. (1990). Comorbidity of mental disorders with alcohol and other drug abuse. Results from the Epidemiologic Catchment Area (ECA) study. *Journal of the American Medical Association*, *264*(19), 2511–2518.

Sabb, F., Burggren, A., Higier, R., Fox, J., He, J., Parker, D., et al. (2009). Challenges in phenotype definition in the whole-genome era: Multivariate models of memory and intelligence. *Neuroscience*, *164*(1), 88–107.

SAS Institute Inc. (2008). *SAS(r) 9.2 enhanced logging facilities*. Cary, NC: SAS Institute Inc.

Saxon, A. J., Oreskovich, M. R., & Brkanac, Z. (2005). Genetic determinants of addiction to opioids and cocaine. *Harvard Review of Psychiatry*, *13*(4), 218–232.

Seber, G. A. F. (1984). *Multivariate observations*. New York: Wiley.

Substance Abuse and Mental Health Services Administration (2007). Results from the 2006 National Survey on Drug Use and Health: National findings. *Office of Applied Studies, NSDUH Series H-32, DHHS Publication No. SMA 07–4293) Rockville*.

Tan, P. -N., Steinbach, M., & Kumar, V. (2006). *Introduction to data mining*. Boston: Pearson Addison Wesley.

Theodoridis, S., & Koutroumbas, K. (1999). *Pattern recognition*. San Diego: Academic Press.

van der Laan, M., Pollard, K., & Bryan, J. (2003). A new partitioning around medoids algorithm. *Journal of Statistical Computation and Simulation*, *73*(8), 575–584.