National Institute on Drug Abuse Center for Genetic Studies

Opioid-Cocaine Dependence Genetics Initiative, Version 2.1 Notes

Summer 2023

Introduction to Opioid-Cocaine 2.1 (July 2023)

Distribution 2.1 of the NIDA Opioid-Cocaine Dependence Genetics Initiative adds n=279 Study 20 subjects to the dataset, resulting in n= 7007 subjects in 4040 families. N=5344 subjects have cell lines. Note that subjects and families have dual ascertainment with either opioid or cocaine dependence. Data for Studies 1 and 20 were managed together, see notes below.

1. Study 20: Nicotine dep/ abuse not assessed
2. Study 20: Stimulant dependence/abuse defined as SCID amphetamine, phentermine dep/abuse.
3. Study 20: clinic data was deposited in two files with different formats, NIDA_SAMPLES and mod3_nida_asi5_scdminicasesonly; demographics and dx var were extracted and harmonized for these two tables; the SCID and ASI data are available separately
4. Study 20: for SCID data in NIDA_SAMPLES.sas7bdat, for subjects with multiple records, the dependence and abuse individual items in the harmonized distribution file were set to the most severe dx in the order Current, Partial Remission and Remission; the derived binary dx were defined as follows: Affected = Current or Partial Rem, Not Affected = Remission; refer to interview dataset for original values.
5. Stimulant abuse/ dep was defined as amphetamine or phentermine (aka "phen"); Other abuse/ dep was defined as hallucinogens, inhalants, PCP, or other.
6. the subject age is set to the min subject age among multiple records; therefore age may not match the dependence/abuse dx in the harmonized table; refer to the full multi-record dataset available in the interview download section;
7. Study 20: this is the dataset indexed by the 7-digit numeric genetics id, instead of the SSADA id (center-family-subject format)
8. Study 20: subject 1716425 is in both clinic interview datasets, ie. both NIDA_SAMPLES and mod3_nida_asi5_scdminicasesonly
9. Study 20: case-control study, fam_id set to ind_id; proband set to NULL
10. Study 20: two records with discrepant sex were removed with genetics IDs = 124018, 4076502; note that 4076502 had a second record with non-discrepant sex that was retained
11. Study 20: the following three interview datasets that are available “as is” from the investigator; note one subject is in both the N=321/ N=325 datasets and the N=6 data; note2: datasets contain duplicate records per subject
    a. N=321 subjects: SCID
    b. N=325 subjects: ASI 5
    c. N=6: ASI 5 (different format than above)
12. Study 20: formats for the interview datasets were output to SAS datasets and may be read with the proc format cntlin option; also saved as csv files; one format dataset/ csv for each of the three datasets
13. Dist 2.1: two subjects were present in both the first Study 20 deposit and the second (current) Study 20 deposit with the same RUID but different subject IDs; the diagnostic

data for the two records were coalesced into one record and the more recent subject ID retained; the two subjects are the following:

a. RUID = 10NA45329, Subject ID1=95-0483-00, Subject ID2=95-3132899 (retained)
b. RUID = NA0048744, Subject ID1=95-0537-00, Subject ID2=95-1346400 (retained)

National Institute on Drug Abuse Center for Genetic Studies

Opioid-Cocaine Dependence Genetics Initiative, Version 2.1 Notes

Summer 2023

Introduction to Opioid-Cocaine 2.0 (Nov 2022)

Distribution 2.0 of the NIDA Opioid-Cocaine Dependence Genetics Initiative adds n=219 Study 20 subjects to the dataset and an additional n=2875 subjects for Study 1, resulting in n= 6730 subjects in 3763 families. N=5067 subjects have cell lines. Note that subjects and families have dual ascertainment with either opioid or cocaine dependence. Data for Studies 1 and 20 were managed together, see notes below.

1. Study 20: two types of subject ids exist for Study 20, the SSADDA id and the genetics id; the SSADDA id indexes the clinic data and the genetics id indexes the genetic dataset; the SSADDA id format is center+family+subject; the genetics id is a 7-digit random integer; the link between SSADDA id and genetics id was available for n=444 subjects.
2. Study 20: the ind_id in the distribution file is the SSADDA id, ie. clinic id; linking ids between the Study 20 genetics id and SSADDA id were obtained and the ind_id was set to the SSADDA_id; all Study 1 subjects have a SSADDA id; therefore, all subjects in Dist 2.0 have a SSADDA id.
3. Study 20: n=39 ids (SSADDA format: center+family+subject) in the RUCDR sample dataset (LIMS) were manually cleaned in coordination with the RUCDR data management team; eg. reformatted with leading zeros and hyphens added to separate (cent, fam, ind) or proband ind code "-00" added; some SSADDA ids were in the genetics id column and some genetics ids were in the SSADDA id column; the most reliable key for this data is the RUID which indexes the blood samples/ cell lines; the subcode field is mostly (>90%) the random genetics id; otherwise subcode is SSADDA id in the remainder of cases;
4. Study 20: merge of clinic with RUCDR LIMS data; 3 records removed with discrepant sex/gender.
5. Dist 2.0: Pedigree structure is missing for version 2 data; families that overlap with version 1 families will have members that are disconnected but share same family id; if genotyped, relationships may be inferred by IBD analysis.
6. Dist 2.0: Study design. Study 1 is a family design while Study 20 is case-control so the partial family structure only applies to Study 1.
7. Dist 2.0: Study 1. Original coding for diagnostic var: 0=Abuse, 1=Unaffected, 2=Dependent, Blank=Not Interviewed; for dependence dx, recodes were 0=Abuse-> blank-missing, 1=Unaffected->Unaffected, 2=Dependent->5=Dependent, Blank=Not Interviewed=same; for abuse dx, recodes were 0=Abuse-> 5=Affected, 1=Unaffected->Unaffected, 2=Dependent->blank-missing, Blank=Not Interviewed=same.
8. Dist 2.0: Version 2 data was merged with Version 1 data for inclusion in the repository; inclusion was the same as Version 1; families were included if they had one or more members with a diagnosis of either opioid dependence or cocaine dependence; family members with no clinic data were excluded, ie. “blood only” subjects; family members with some clinic data were included.
9. Dist 2.0: removed n=9 subjects with discrepant sex between v1 and v2.
10. Dist 2.0: n=19 subjects/ families with no probands because proband was defined as subject id = “00” and the subject/ family was included because a relative other than the proband was affected with blood.

11. Dist 2.0: n=21 subjects with no sex or phenotype data but *with* bloods from Study 1 were included because no pedigree data was available and these subjects could possibly be informative genetically

Introduction to Opioid-Cocaine 1.0 (Aug 2019)

Distribution 1.0 of the NIDA Opioid-Cocaine Dependence Genetics Initiative contains 3636 subjects in 793 families from Study 1. N=1973 subjects have cell lines. Due to dual ascertainment of families with either opioid or cocaine dependence, Study 1 was removed from Opioid Distribution 5.5 and the new Opioid-Cocaine distribution was formed.

Notes

1. Distribution file. Study 1. Original coding for diagnostic var: 0=Abuse, 1=Unaffected, 2=Dependent, Blank=Not Interviewed; for dependence dx, recodes were 0=Abuse-> blank-missing, 1=Unaffected->Unaffected, 2=Dependent->5=Dependent, Blank=Not Interviewed=same; for abuse dx, recodes were 0=Abuse-> 5=Affected, 1=Unaffected->Unaffected, 2=Dependent->blank-missing, Blank=Not Interviewed=same.
2. Distribution file. Study 1. families were included if they had one or more members with a diagnosis of either opioid dependence or cocaine dependence. Family members with no pedigree connectors, ie. father, mother id, and not a founder, were excluded, even if they had blood samples and some demographic information (sex, race); typically if connecting ids were missing, diagnostic data was also missing.
3. Distribution file. Study 1. Sex of two subjects was set to missing due to discrepant reports at different time points: 10-9044-00, 12-4200-00.