12-21-2017 (prepared by Melanie Wall, Keely Cheslack-Postava, and Tianshu Feng in the Department of Psychiatry at Columbia University Medical Center and the Research Foundation for Mental Hygiene)

Supplementary Programming Methods for Analysis of Drug Initiation Sequences in Wall MM, Cheslack-Postava K, Hu MC, Feng T, Griesler P, Kandel DB. Nonmedical prescription opioids and pathways of drug involvement in the US: Generational differences, *Drug and Alcohol Dependence*, 182: 103-111. 2017.

This document describes and provides the statistical programming code (using SAS and R software) and the input data and snapshots of output data for the simulation-based method developed and implemented for testing the statistical significance of specific drug initiation sequences in the Wall et al. (2017) *DAD* paper.

Step 1. Obtain NSDUH data and re-code key variables.
Public use data files for the NSDUH are available at
https://www.samhsa.gov/samhsa-data-outcomes-quality/major-data-collections/public-use-files-2014-nsduh
OR
https://www.icpsr.umich.edu/icpsrweb/ICPSR/series/64

Input: Public use data files for 2013 and 2014
Program: Sequences Step 1.sas. Variables for age at interview (ageyrs), and ever use (1/2 = no/yes) and the age at initiation for 6 substances (alcohol (alc), cigarettes (cig), marijuana (mj), cocaine (coc), opiates (analgesics anl), and heroin (her)) were recoded. The variable "year" was created to correspond to each year of downloaded data. The category variable "age5" was created corresponding to 3 generations included in the study, as well as the younger and older subjects.
Output: SAS data set nsduh_2013_2014 (first 10 observations shown below; n total=110,431).

Obs	AGE2	year	id	mjever	cigever	alcever	anlever	cocever	herever	cig_age	alc_age	mj_age	anl_age	coc_age	her_age	ageyrs	age5	weight2
1	9	2013	48694667	1	2	1	1	1	1	19	.	.	.	.	.	20.0	2	2199.20
2	7	2013	88530883	2	2	2	1	2	1	14	13	14	.	16	.	18.0	2	709.60
3	6	2013	33251077	1	2	2	1	1	1	14	16	.	.	.	.	17.0	1	7026.31
4	17	2013	37814127	1	2	2	1	1	1	16	18	.	.	.	.	70.0	5	5424.09
5	15	2013	18762590	1	2	2	1	1	1	14	18	.	.	.	.	42.5	3	2825.87
6	14	2013	15866325	1	1	2	1	1	1	.	18	.	.	.	.	32.0	2	2348.51
7	17	2013	19783709	1	2	2	1	1	1	22	21	.	.	.	.	70.0	5	16585.47
8	12	2013	35106150	1	2	2	1	1	1	22	19	.	.	.	.	24.5	2	1605.62
9	8	2013	67182690	2	2	2	2	2	1	16	14	15	17	18	.	19.0	2	3198.37
10	2	2013	68185320	1	1	1	1	1	1	.	.	.	.	.	.	13.0	1	1191.13

Step 2. Create summaries of NSDUH population to be used as input parameters in simulation
In Step 3 we will simulate a population with characteristics identical to the observed NSDUH sample (including ages at survey, prevalences of lifetime use, distributions of onset age for each drug). To do this, in Step 2 we must obtain the population values for correlations between lifetime use of different drugs and frequencies of ages of onset among lifetime users of each drug. These population values will be taken to be the observed values from the NSDUH sample for each separate age at time of survey. To address small sample sizes in some cells which result in zero prevalences, a moving average method was used to smooth prevalences across age groups by combining age with age plus or minus one.

Input: SAS data set "nsduh_2013_2014": NSDUH data from 2013 and 2014, processed as described in Step 1

Program: Sequences Step 2.sas – This program produces 4 output files with parameters that will be used to simulate data in Step 3.
Output:

a. N (sample size for simulation) for each age group. Age from public NSDUH data is in single years from 12-21, then grouped as 22-23, 24-25, 26-29, 30-34, 35-49, 50-64 and >=65. **Output:** Appendix O1.csv
b. Polychoric correlation matrix **Output:** Appendix O2.csv
c. Probability of ever use **Output:** Appendix O3.csv
d. Age of onset frequency among lifetime users. Note, probability of use at each onset age, up to and including the survey age add up to 100%. **Output:** Appendix O4.csv

Step 3. Simulate data with sequential patterns of drug use initiation generated randomly: Using the parameters in Appendix O1-O4 from Step 2, use (1) a multivariate probit model to simulate correlated dichotomous lifetime use of all 6 drugs, (2) a multinomial distribution to simulate onset age for each drug, conditional on lifetime use.

Input: Appendix O1.csv, Appendix O2.csv, Appendix O3.csv, and Appendix O4.csv
Program: Sequences Step 3.R – This program simulates 10 random samples with size and age distribution equal to the full NSDUH and with drug use initiation generated randomly from a common liability model where drugs are correlated but not directly influential on one another.
Output: Simulated data.CSV – a CSV file containing variables for age at survey, and for each of 6 drugs (alcohol, cigarettes, marijuana, cocaine, opiates, heroin) – a) a dichotomous variable indicating ever use, and b) a variable for age at first use (missing if not used). This file includes N= 1,103,230 which is 10x the total number of observations as the original NSDUH data set on which it is based. (This number is slightly lower than that from step 1 due to the exclusion of observations with reported use of marijuana, cocaine, opiates, or heroin before age 5.)

	ageyrs	alcever	anlever	cigever	cocever	herever	mjever	alc_age	anl_age	cig_age	coc_age	her_age	mj_age
1	12	0	0	0	0	0	0						
2	12	0	0	0	0	0	0						
3	12	0	0	0	0	0	0						
4	12	0	0	0	0	0	0						
5	12	0	0	0	0	0	0						
6	12	1	0	1	0	0	0	9		10			
7	12	0	0	0	0	0	0						
8	12	0	0	0	0	0	0						
9	12	0	0	0	0	0	0						
10	12	0	0	0	0	0	0						

Step 4. Prepare simulated data and original NSDUH data for analysis of initiation sequences

a. For simulated data, index the data set and attach weights
The simulated data from Step 3 includes10 population samples with identical age at time of survey distribution to the original NSDUH 2013-2014 sample. Each simulated observation is given an identification number (id), assigned a sampling weight (weight2) corresponding to the original NSDUH sample (based on their age), and indexed from 1-10 (dset) corresponding to the 10 simulations.

Input: Simulated data.CSV (output from Step 3)
Program: Sequences Step 4a.sas
Output: SAS data set simdata2 – includes the variables from the input data set, and variables for: data set #, an assigned ID#, weight, year, and additional age variables (first 10 observations shown below).

SAS Data set SIMDATA2, first 10 observations

Obs	agevrs	alcever	anlever	cigever	cocever	herever	mjever	alc_age	anl_age	cig_age	coc_age	her_age	mj_age	dset	ID	AGE2	VESTR	VEREP	year	weight2	age5
1	12	0	0	0	0	0	0	.		.				1	1	1	30031	02	2013	187.44	1
2	12	0	0	0	0	0	0	.		.				2	1	1	30031	02	2013	187.44	1
3	12	0	0	0	0	0	0	.		.				3	1	1	30031	02	2013	187.44	1
4	12	0	0	0	0	0	0	.		.				4	1	1	30031	02	2013	187.44	1
5	12	0	0	0	0	0	0	.		.				5	1	1	30031	02	2013	187.44	1
6	12	1	0	1	0	0	0	9		10				6	1	1	30031	02	2013	187.44	1
7	12	0	0	0	0	0	0	.		.				7	1	1	30031	02	2013	187.44	1
8	12	0	0	0	0	0	0	.		.				8	1	1	30031	02	2013	187.44	1
9	12	0	0	0	0	0	0	.		.				9	1	1	30031	02	2013	187.44	1
10	12	0	0	0	0	0	0	.		.				10	1	1	30031	02	2013	187.44	1

b. For simulated and original NSDUH data, separately:

- **Reduce to 5 drugs:** For simplicity of presentation, initiation of alcohol and cigarettes were combined as if they were a single drug with age at initiation set to age of first use of either alcohol or cigarettes
- **Handle ties in age of initiation:** When a person reports they started using more than one drug at the exact same age, this creates a problem for defining sequences which require a strict order. To accommodate this phenomena of a "tie" in the age of onset, whenever a "tie" is observed, we break it by randomly assigning the order of the sequences based on the probability distribution of observed sequences for tie-involved drugs (see separate **Tie Supplement** for a more complete description of the procedure). Note, this step involves a large amount of coding

Inputs: SAS data sets simdata2 (simulated data, step 4a); nsduh_2013_2014 (actual data, step 1) are separately processed.
Program: Sequences Step 4b.sas
Outputs: SAS data sets "simdata_filled_ties" and "nsduh_filled_ties". (Below are 25 observations from the "Baby Boomer" (AGE5=4) group from simdata_filled_ties shown below).
Note: The ties column summarizes whether there were any age of onset ties between the 5 drug onset ages (B_age (minimum of cig_age (cigarette) and alc_age (alcohol)), M_age = mj_age (marijuana), O_age = anl_age (non-medical prescription opioids), C_age = coc_age (cocaine), H_age = her_age (heroin)). For example, ties= 11111 is no ties. Ties=13331 is a 3-way tie for the 2nd through 4th drugs initiated. ties=22333 means the 1st and 2nd drugs were used at the same age, and the 3rd-5th were used at an age that is equal, but higher than that for the first 2 drugs. The columns dr_1-dr_5 and dr_seq are created after randomly breaking the ties within the program.

Obs	ageyrs	alcever	anlever	cigever	cocever	herever	mjever	alc_age	anl_age	cig_age	coc_age	her_age	mj_age	dset	AGE2	VESTR	VEREP	year	weight2	age5	ID	S_age	A_age	M_age	O_age	C_age	H_age	B_age	dr_1	dr_2	dr_3	dr_4	dr_5	dr_seq	ties
1	57.5	1	0	1	0	0	0	18	.	13		.		1	16	40049	01	2014	7219.16	4	100000_1	13	18	.	.	.	.	13	B	_	_	_	_	B___	11111
2	57.5	1	0	0	1	0	1	23	.	.	24	.	21	10	16	40049	01	2014	7219.16	4	100000_10	.	23	21	.	24	.	23	M	B	C	_	_	MBC_	11111
3	57.5	1	0	0	0	0	0	3	.	.	.	.	.	2	16	40049	01	2014	7219.16	4	100000_2	.	3	.	.	.	.	3	B	_	_	_	_	B___	11111
4	57.5	1	1	1	0	0	1	18	18	14	18	.	21	3	16	40049	01	2014	7219.16	4	100000_3	14	18	21	18	18	.	14	B	C	O	M	_	BCOM_	12211
5	57.5	1	0	1	0	0	1	16	.	20	.	.	22	4	16	40049	01	2014	7219.16	4	100000_4	20	16	22	.	.	.	16	B	M	_	_	_	BM___	11111
6	57.5	1	1	1	1	0	1	22	32	17	24	.	30	5	16	40049	01	2014	7219.16	4	100000_5	17	22	30	32	24	.	17	B	C	M	O	_	BCMO_	11111
7	57.5	1	0	1	1	0	1	21	.	11	16	.	21	6	16	40049	01	2014	7219.16	4	100000_6	11	21	21	.	16	.	11	B	C	M	_	_	BCM__	11111
8	57.5	1	0	1	0	0	1	8	.	15	.	.	16	7	16	40049	01	2014	7219.16	4	100000_7	15	8	16	.	.	.	8	B	M	_	_	_	BM___	11111
9	57.5	1	0	1	0	0	1	16	.	13	.	.	18	8	16	40049	01	2014	7219.16	4	100000_8	13	16	18	.	.	.	13	B	M	_	_	_	BM___	11111
10	57.5	1	0	0	0	0	1	35	.	.	.	.	21	9	16	40049	01	2014	7219.16	4	100000_9	.	35	21	.	.	.	35	M	B	_	_	_	MB___	11111
11	57.5	1	1	1	1	0	1	19	17	14	17	.	13	1	16	40020	02	2014	4910.63	4	100001_1	14	19	13	17	17	.	14	M	B	O	C	_	MBOC_	11221
12	57.5	1	0	1	0	0	0	19	.	13	.	.	.	10	16	40020	02	2014	4910.63	4	100001_10	13	19	.	.	.	.	13	B	_	_	_	_	B___	11111
13	57.5	1	0	1	1	0	1	18	.	12	23	.	17	2	16	40020	02	2014	4910.63	4	100001_2	12	18	17	.	23	.	12	B	M	C	_	_	BMC_	11111
14	57.5	1	1	1	1	0	1	14	25	7	25	.	18	3	16	40020	02	2014	4910.63	4	100001_3	7	14	18	25	25	.	7	B	M	C	O	_	BMCO_	11221
15	57.5	1	0	1	0	0	1	11	.	17	.	.	17	4	16	40020	02	2014	4910.63	4	100001_4	17	11	17	.	.	.	11	B	M	_	_	_	BM___	11111
16	57.5	1	0	1	0	0	0	16	.	18	.	.	.	5	16	40020	02	2014	4910.63	4	100001_5	18	16	.	.	.	.	16	B	_	_	_	_	B___	11111
17	57.5	1	0	1	0	0	1	17	.	14	.	.	16	6	16	40020	02	2014	4910.63	4	100001_6	14	17	16	.	.	.	14	B	M	_	_	_	BM___	11111
18	57.5	1	0	1	0	0	1	16	.	18	.	.	16	7	16	40020	02	2014	4910.63	4	100001_7	18	16	16	.	.	.	16	M	B	_	_	_	MB___	22111
19	57.5	1	0	1	0	0	1	20	.	13	.	.	19	8	16	40020	02	2014	4910.63	4	100001_8	13	20	19	.	.	.	13	B	M	_	_	_	BM___	11111
20	57.5	0	0	0	0	0	0	.	.	.	.	.	.	9	16	40020	02	2014	4910.63	4	100001_9	.	.	.	.	.	.	.	_	_	_	_	_	_____	11111
21	57.5	1	0	1	0	0	1	20	.	15	.	.	14	1	16	40045	01	2014	4784.97	4	100002_1	15	20	14	.	.	.	15	M	B	_	_	_	MB___	11111
22	57.5	1	0	1	0	0	1	16	.	15	.	.	17	10	16	40045	01	2014	4784.97	4	100002_10	15	16	17	.	.	.	15	B	M	_	_	_	BM___	11111
23	57.5	1	0	1	0	0	1	16	.	12	.	.	22	2	16	40045	01	2014	4784.97	4	100002_2	12	16	22	.	.	.	12	B	M	_	_	_	BM___	11111
24	57.5	1	0	1	0	0	0	23	.	14	.	.	.	3	16	40045	01	2014	4784.97	4	100002_3	14	23	.	.	.	.	14	B	_	_	_	_	B___	11111
25	57.5	1	0	1	0	0	0	12	.	16	.	.	.	4	16	40045	01	2014	4784.97	4	100002_4	16	12	.	.	.	.	12	B	_	_	_	_	B___	11111

Step 5. Calculation of the expected prevalence of drug initiation sequences using the simulated population data
Using the simulated sequences of initiation of 5 drugs processed in step 4, for each generation age group (Millenial, GenX, Baby Boomer) estimate the weighted prevalence of each drug initiation sequence (using the NSDUH sampling weights). Repeat for all 10 simulated datasets and take the average prevalence for each sequence across all ten datasets. These averages constitute the 'expected' prevalence for each drug initiation sequence for that generation age group.

Input: SAS data set simdata_filled_ties.
Program: Sequences Step 5.sas
Output: SAS data set Exp_seqprev – Lists drug sequence and the expected prevalence as well as the standard deviation of those prevalences (across the 10 simulated data set). The first 10 sequences are shown below for Millennials.

Obs	dr_seq	per_exp_avg	per_exp_std
1	B____	32.1968	0.36937
2	BM___	20.3891	0.36216
3	MB___	10.0641	0.25079
4	_____	9.7345	0.20983
5	BMO__	3.7453	0.17346
6	BMC__	2.4872	0.13074
7	BO___	2.0424	0.06394
8	MBO__	1.9717	0.11812
9	BOM__	1.8059	0.04595
10	BMCO_	1.3239	0.05865

Step 6. Calculation of the observed prevalence of drug initiation sequences in the NSDUH
Using the observed NSDUH data obtained in Step 1 and processed in Step 4b, for each generation age group (Millenial, GenX, Baby Boomer) estimate the weighted prevalences of each drug initiation sequence.

Input: SAS data set nsduh_filled_ties.
Program: Sequences Step 6.sas
Output: SAS data set Obs_seqprev – Lists drug sequences, the weighted prevalence of the sequence in the observed data, and the raw count for the sequence (unweighted) in the observed data. The first 10 sequences are shown below for Millennials.

Obs	dr_seq	percent_obs	rawN_obs
1	B____	32.2087	14451
2	BM___	25.1796	11271
3	_____	9.9817	4815
4	BMO__	5.8583	2732
5	MB___	5.1019	2560
6	BMC__	4.1856	1616
7	BMOC_	3.2182	1399
8	BMCO_	2.3362	866
9	BO___	1.8755	829
10	MBO__	1.4079	700

Step 7. Comparison of the observed and expected prevalence of drug initiation sequences, and calculation of statistical significance
Compare the observed and the expected frequencies of each drug initiation sequence using a standard test of proportions, with p-values based on the standard normal distribution and Bonferonni adjustment for the total number of sequences tested.

Inputs: SAS data sets Obs_seqprev (created in Step 6, above) and Exp_seqprev (created in step 5, above)
Program: Sequences Step 7.sas
Output: SAS data set fin_all (10 sequences shown below; shown for Millennials). Includes for each drug sequence: the observed and expected prevalence; the z-score and p-value from the test of proportions; and an indicator for statistical significance using a Bonferonni adjustment to alpha=0.05. The variable "totalpop" indicates the number of observations in the population stratum for which the test was applied. **These values are found in Supplemental Table 1 of the Wall et al 2017 manuscript in DAD and form the primary tests for determining significant sequences.**

Obs	dr_seq	pvalue	z	totalpop	p_sig_bonferonni	obs	exp
1	BMOC_	0.0000000000000	40.6575	45010	1	3.2182	1.1639
2	BMOCH	0.0000000000000	31.9656	45010	1	0.7239	0.1470
3	BM___	0.0000000000000	25.2222	45010	1	25.1796	20.3891
4	BMO__	0.0000000000000	23.6143	45010	1	5.8583	3.7453
5	BMC__	0.0000000000000	23.1382	45010	1	4.1856	2.4872
6	BMCOH	0.0000000000000	20.2506	45010	1	0.5635	0.1705
7	BMCO_	0.0000000000000	18.8101	45010	1	2.3362	1.3239
8	OCBH_	0.0000000000000	11.8059	45010	1	0.0039	0.0001
9	BMCH_	0.0000000000000	11.8046	45010	1	0.2169	0.0703
10	MBOCH	0.0000000000186	6.7165	45010	1	0.1585	0.0725