

NSDUH Drug Initiation Sequences – Updated methods to address missing data resulting from ties

The results presented here pertain to the observed data (i.e. not simulated), from NSDUH years 2011-2014. Five drugs are included: B “both” refers to cigarettes/alcohol, M=marijuana, C=cocaine, O=opiates, and H=heroin. Observations where initiation of M, C, O, or H was reported at ages <5 were excluded for all analyses.

Drug initiation sequences are defined for each individual based on the age of first use for each of these 5 drugs that they report having ever used. When a respondent reports initiating 2 or more of these drugs at the same age, the sequence cannot be uniquely determined and missing data results.

In the first version of this method, this missing data was addressed by assuming that where the sequence was unknown (tied), the drugs were always initiated in decreasing order of prevalence of use in the total population. However, one problem with this method was that it resulted in many sequences that arose solely due to ties, and did not occur in the data in individuals without missing data.

In the updated version, we now address ties based on actual order of drug initiation among observations without missing information (i.e. ties).

- 2, 3, and 4-way ties were addressed in the following manner:
 1. For every possible combination of k=2, 3, or 4 drugs, all observations where each of those drugs was used, beginning at different ages are identified in the data.
 2. Then, the prevalence of each permutation of the k drugs among observations where they were initiated at different ages is calculated, incorporating sampling weights. For k=2 drugs, the prevalences are calculated separately within 5 different age groups; for 3 and 4 drugs, all ages are combined.
 3. Next, for every k-way tie, one of the permutations of drugs for the relevant drug combination is assigned, based on a random draw from a probability distribution based on the table of observed prevalences (or binomial for 2-way ties).
 4. Currently, we perform only one imputation.
 5. Finally, for every combination, the set of imputed orders is compared to the observed orders to assess the performance of the process (i.e. make sure the code worked right).
- 5-way ties are very infrequent in the data (5 out of over 200,000 observations and there are 120 permutations, only 29 of which actually occurred. Of these 29, the 3 most frequent constituted over 70% of observations where all 5 drugs were used at different ages. To simplify the process, 5-way ties were assigned to one of these top 3 sequences based on a random draw from a distribution with equal probability.

The results below show some description of ties in the data, non-tied sequences which were used to inform probabilities of assigning ties to orders, evaluation of the implemented process, and basic results from the observed data after imputing missing sequence data.

Observed ties. Ties occurred in 12.5% of all observations (Table 1), with the vast majority being 2-way ties.

Table 1. Count and percentage (unweighted) of N=224,096 total observations with any tie and with different types of ties.

Type of tie	N	%
Any tie	28,000	12.49
2-way (any)	26,552	11.85
3-way	1403	0.63
4-way	93	0.04
5-way	5	0.00
2 x 2-way	874	0.39
2-way + 3-way	53	0.02

Observed prevalences of initiation orders based on observations where the relevant drugs were all used, but initiated at different ages. Percentages were calculated incorporating sampling weights.

Table 2. 2-way ties.

Drugs tied	Count of obs. with this type of tie	Percent of non-tied observations that used both drugs where the first drug listed in column 1 was used first, by age group*				
	N (total=27,426)	Age 12-17	Age 18-25	Age 26-34	Age 35-49	Age 50+
BC	238	97.4	99.1	99.4	99.4	99.7
BH	25	93.8	99.0	99.7	99.7	99.4
BM	19171	74.8	80.2	87.0	90.3	93.7
BO	1518	79.1	90.8	95.8	96.2	96.6
CH	417	69.8	87.1	90.0	91.9	71.9
MC	1299	96.7	98.0	97.5	96.8	98.2
MH	42	94.8	98.4	98.5	99.2	98.4
MO	2376	74.6	86.3	90.9	93.2	91.0
OC	2140	79.3	72.1	51.6	30.0	40.2
OH	200	80.7	92.1	83.7	67.3	60.5

*i.e. where the drugs tied is noted as "BC", 97.4% of 12-17 year olds who reported using alcohol/cigarettes (B) and cocaine (C) at different ages had used B first.

Table 3. 3-way ties.

Drugs tied	Count of obs. with this type of tie	Percentage of non-tied observations that used the three drugs in each possible order*					
	N (total=1403)	p_123	p_132	p_213	p_231	p_312	p_321
BMC	313	85.90	2.56	11.17	0.25	0.10	0.02
BMO	700	77.64	8.08	11.63	0.49	1.79	0.37
BMH	15	85.69	1.43	12.39	0.29	0.06	0.14
BOC	20	42.90	55.23	1.45	0.11	0.29	0.02
BCH	2	82.83	16.26	0.70	0.01	0.20	0.00
BOH	7	72.05	26.25	1.63	0.00	0.06	0.01
MOC	172	38.32	55.69	4.47	0.26	1.06	0.20
MOH	12	67.10	26.15	6.15	0.11	0.42	0.06
MCH	31	82.80	14.72	1.25	0.08	0.95	0.20
OCH	131	37.17	6.25	31.55	18.47	1.95	4.61

*Order of the 3 drugs based on the positions listed in the first column. Example: for drugs BMC, p_123=85.9, meaning that 85.9% of subjects who used B, M, and C without ties used them in the order B, M, C. p_213=11.2% for BMC means that 11.2% of such subjects used the 3 drugs in the order M, B, C.

Table 4. 4-way ties.

Drugs tied	Count of obs. with this type of tie N (total=93)	Percentage of non-tied observations that used the four drugs in each possible order*															
		p_1234	p_1243	p_1324	p_1342	p_1423	p_1432	p_2134	p_2143	p_2314	p_2341	p_2413	p_2431	p_3124	p_3142	p_3214	p_3241
BMOC	58	32	50	4	0	1	0	5	7	0	0	0	1	0	0	0	0
BMOH	2	57	24	6	0	0	0	8	3	1	0	0	0	0	0	0	0
BMCH	11	70	15	1	0	1	0	11	2	0	0	0	0	0	0	0	0
BOCH	4	34	6	32	19	2	5	1	0	0	0	0	0	0	0	0	0
MOCH	18	31	4	33	20	1	4	4	2	0	0	0	1	0	0	0	0

*Order of the 4 drugs based on the positions listed in the first column. Example: for drugs BMOC, p_1243=50, meaning that 50% of subjects who used B, M, O, and C without ties used them in the order B, M, C, O.

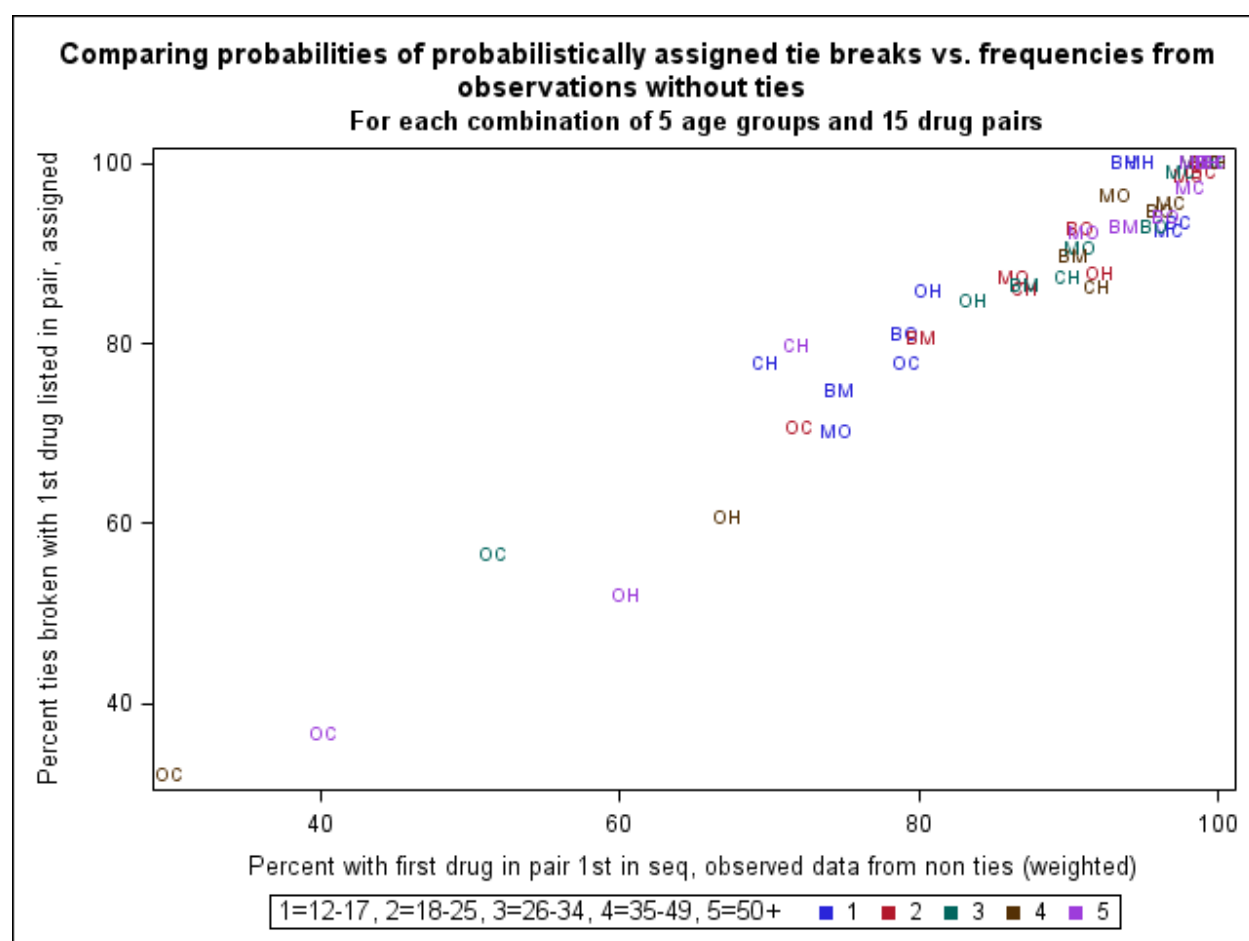
Percentages have been rounded to facilitate display. Shaded (blue) cells indicate observed prevalence >0%. Non-shaded cells indicate sequence orders that were not observed.

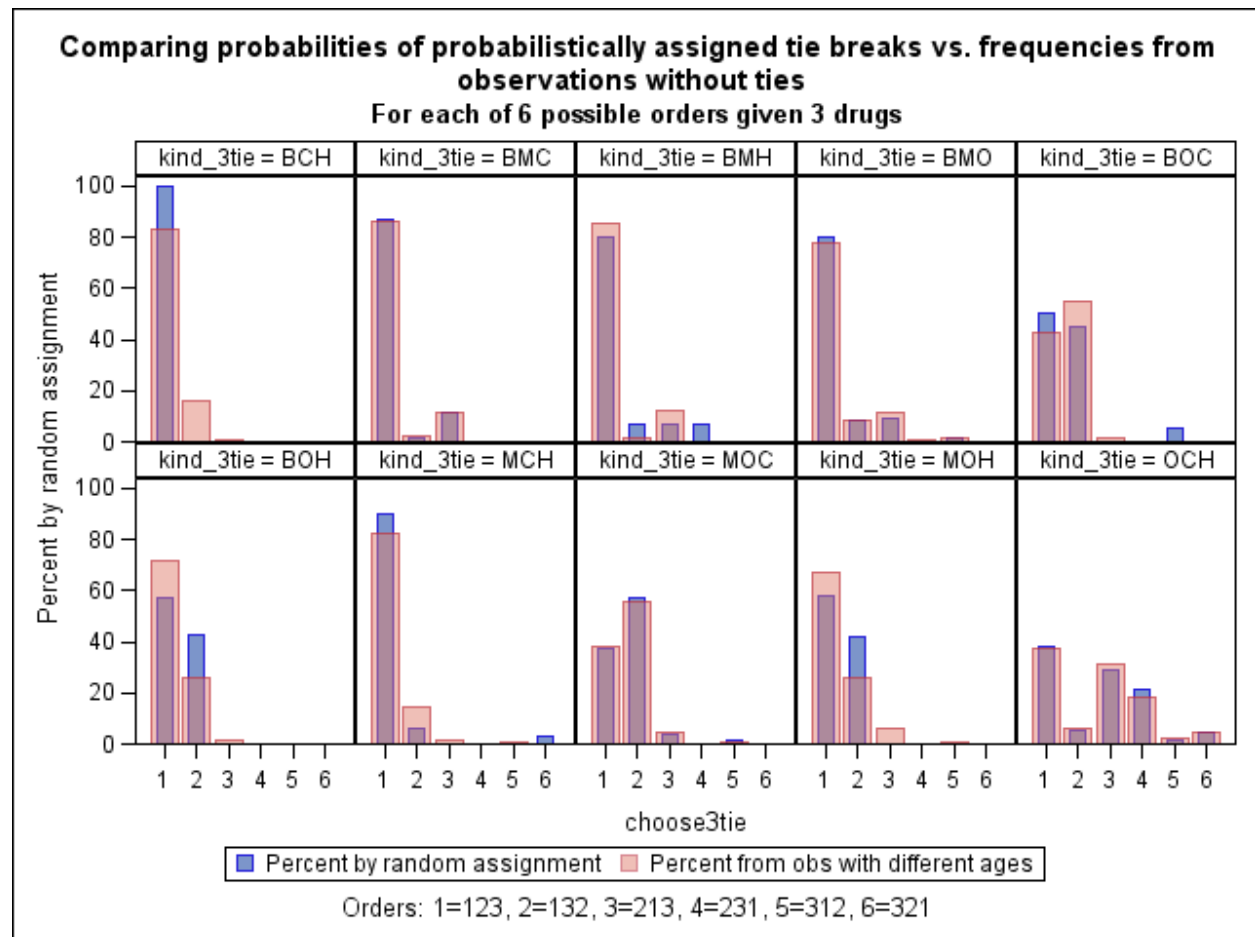
Table 5. 5-way ties.

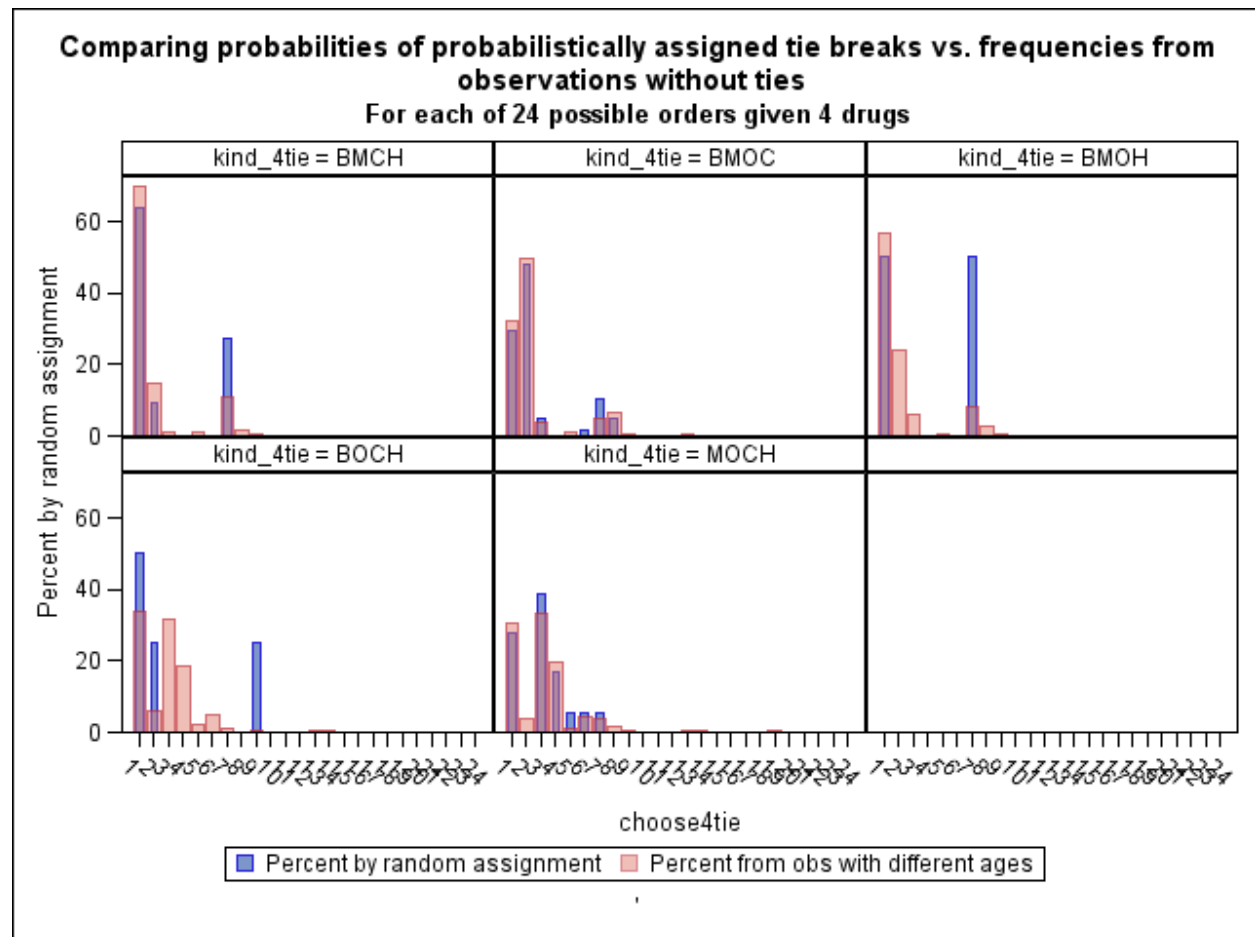
All 5-way ties involve all 5 drugs. A total of only 5 5-way ties were observed in this data. Among observations where subjects reported using all 5 drugs initiated at different ages, 29 out of a possible 120 actual orders occurred.

5-drug order	Percentage of non-tied observations that used the five drugs in this order
BMCOH	29.19
BMOCH	25.27
BMCHO	17.93
BMHCO	4.44
MBOCH	4.42
BMOHC	4.01
MBCOH	3.84
BOMCH	2.78
BOMHC	2.09
MBCHO	1.93
BMHOC	0.79
MOBCH	0.59
BHMOC	0.46
MOCBH	0.35
BOCMH	0.35
BCMOH	0.32
MBHCO	0.32
MBOHC	0.26
MCBHO	0.21
OBMCH	0.14
BCOHM	0.11
OMBCH	0.07
BHOMC	0.04
MBHOC	0.03
BCMHO	0.02
BOCHM	0.02
OBCMH	0.01
BHMCO	0.01
BOHCM	0.01

Assessing performance of the random assignment procedure (2, 3, and 4-way ties only).







Complete list of observed drug sequences following updated procedure for addressing ties, NSDUH 2011-2014.

158 different sequences were observed in the total data.

Drug sequence	Observed percentage*	N*
B__	39.50	69479
BM__	21.34	43502
____	14.23	56111
BMC__	6.57	8929
BMO__	3.32	9084
MB__	2.60	8215
BMCO__	2.29	3759
BO__	1.86	3875
BMOC__	1.76	4247
MBC__	0.77	1280
MBO__	0.57	1990
BOM__	0.54	1579
O__	0.45	1760
OB__	0.37	1252
BMOCH	0.33	828
BMCH__	0.32	395
BMCOH	0.31	586
BC__	0.28	401
MBCO__	0.28	560
MBOC__	0.26	822
M__	0.23	1142
BCM__	0.22	298
BMCHO	0.19	283
BOMC__	0.18	454
OBM__	0.17	602
BMHC__	0.07	78
BMOHC	0.06	142
MBOCH	0.06	155
MOB__	0.05	250
BMH__	0.05	70
BCMO__	0.05	77
MBCOH	0.05	115
BMHCO	0.05	56
BOMCH	0.04	98
MBCH__	0.04	63
OMB__	0.04	169
OBMC__	0.04	109
BCO__	0.04	49
MCB__	0.03	64
BMOH__	0.03	85
BMHOC	0.03	44
MOBC__	0.03	63
MBCHO	0.02	33
BOC__	0.02	43

*Percentages incorporate sampling weights. N are unweighted.

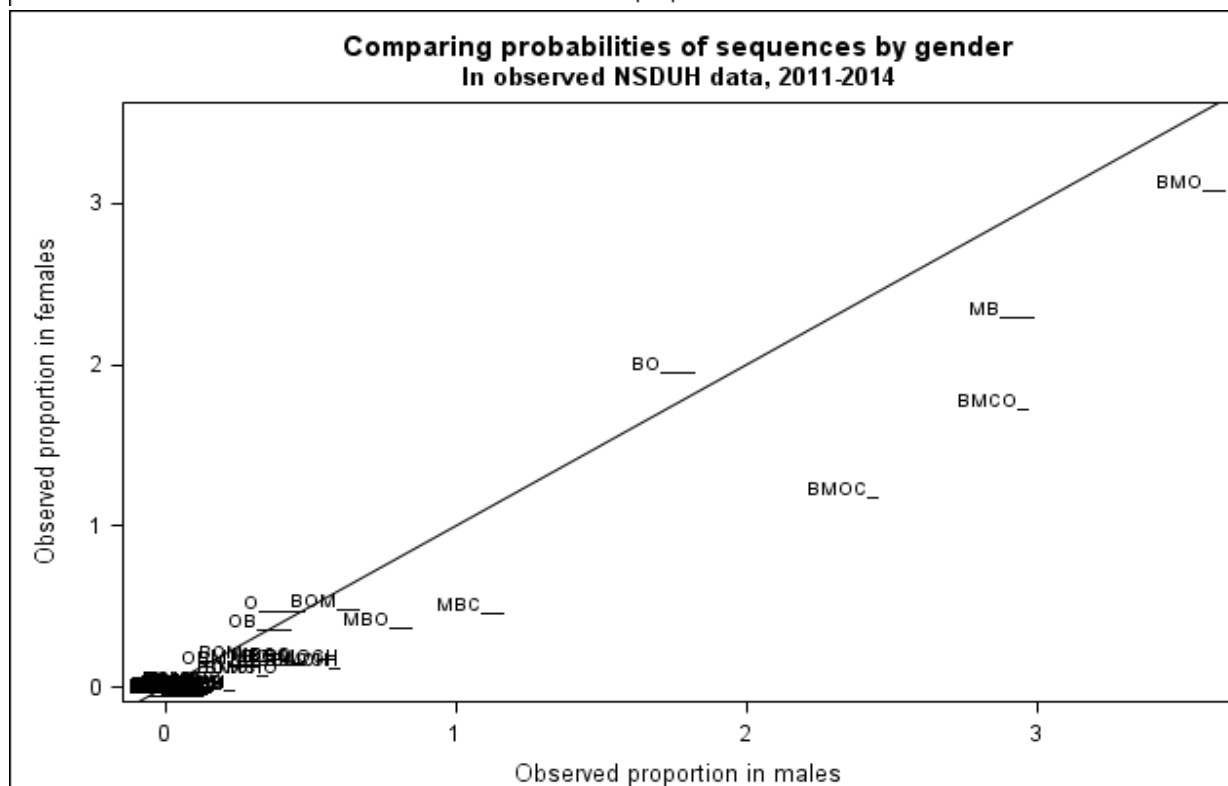
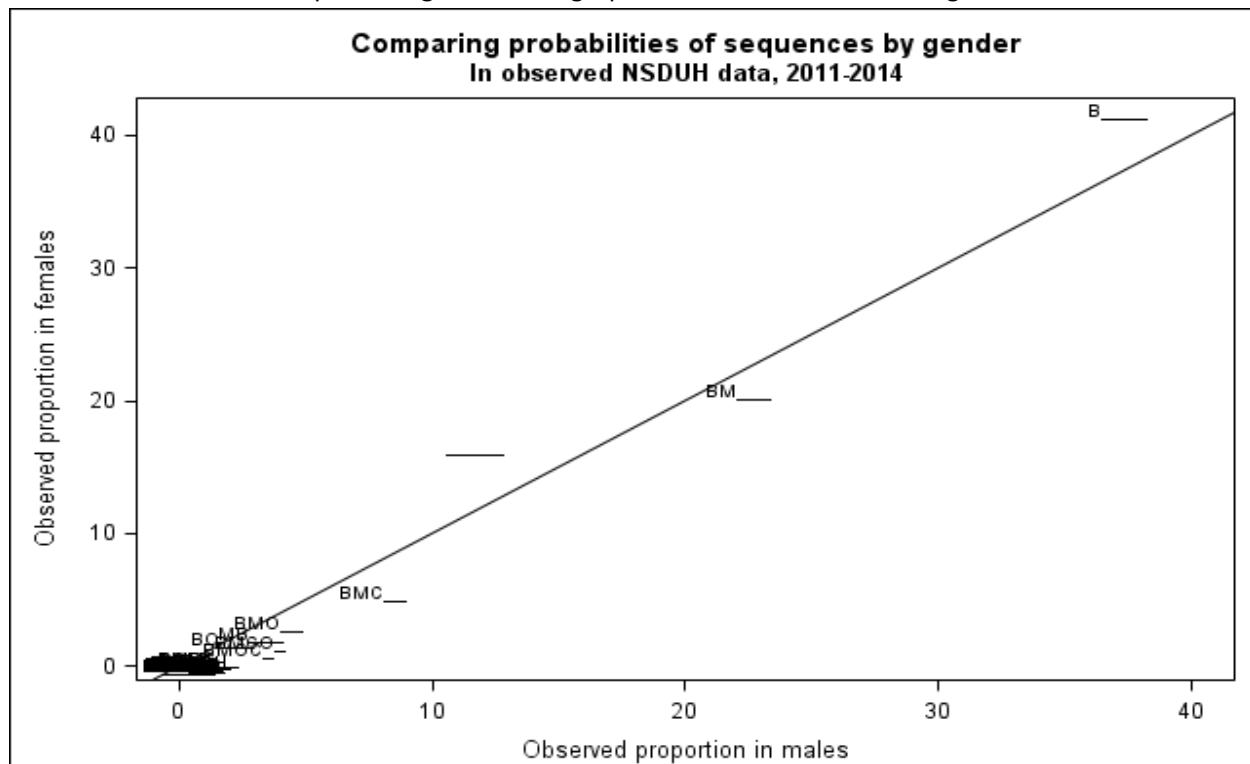
BCOM_	0.02	21
MBHC_	0.01	14
CBM_	0.01	26
BMHO_	0.01	23
BOMHC	0.01	13
MO_	0.01	35
BCH_	0.01	9
OM_	0.01	44
C_	0.01	18
OBMCH	0.01	18
BHMC_	0.01	9
BOCM_	0.01	25
MC_	0.01	14
MBOHC	0.01	19
BCMOH	0.01	13
OMBC_	0.01	26
OBCM_	0.01	11
MCBO_	0.0049	20
MBHOC	0.0045	12
MBHCO	0.0045	8
BCMh_	0.0043	5
CB_	0.0041	10
BHO_	0.0039	6
MBH_	0.0038	15
BHMOC	0.0037	7
MOBCH	0.0035	15
CBMO_	0.0034	7
MBOH_	0.0033	17
BOCMH	0.0032	8
OBC_	0.0032	9
BOMH_	0.0030	11
BHC_	0.0028	5
CMB_	0.0026	3
CBMOH	0.0025	4
BHCM_	0.0024	4
BOCH_	0.0023	6
HB_	0.0022	5
BCHMO	0.0022	3
MCO_	0.0022	2
CBMHO	0.0021	2
MOCB_	0.0020	8
BOH_	0.0020	9
MOCBH	0.0020	4
HMBC_	0.0020	2
BH_	0.0020	13
BHM_	0.0018	4
CH_	0.0017	1
BOHC_	0.0016	2
MH_	0.0015	2
MHB_	0.0014	3
OBMH_	0.0014	5

OMBH_	0.0014	1
MBHO_	0.0014	4
MCBHO	0.0012	4
OMCB_	0.0011	2
OBMHC	0.0011	4
CMBOH	0.0011	3
CBOM_	0.0010	2
BOHMC	0.0009	5
MHBC_	0.0009	1
BCOH_	0.0009	3
MCOH_	0.0009	1
MCHB_	0.0009	3
BHMO_	0.0008	2
MCHOB	0.0007	2
OC__	0.0007	2
BCHO_	0.0006	2
MCOB_	0.0006	2
OCBH_	0.0005	1
BHMC O	0.0005	3
BCOHM	0.0005	1
CMBO_	0.0005	3
OBH_	0.0005	3
OMBCH	0.0005	5
MOHBC	0.0004	1
CMO__	0.0004	1
MHBO_	0.0004	1
COBM_	0.0004	1
HBCMO	0.0004	1
OCBM_	0.0003	2
MOBHC	0.0003	2
MCBOH	0.0003	1
CMOB_	0.0002	2
HCBOM	0.0002	1
H__	0.0002	3
MCHBO	0.0002	1
OBCH_	0.0002	1
COMB_	0.0002	1
BCHOM	0.0002	1
MCBH_	0.0002	2
MOC__	0.0002	1
BCMHO	0.0002	2
BHOMC	0.0002	1
BOCHM	0.0002	2
HO__	0.0002	1
OBHMC	0.0002	1
CO__	0.0001	2
MOBH_	0.0001	1
OMCBH	0.0001	2
CMHB_	0.0001	1
HBM__	0.0001	1
BHCMO	0.0001	1

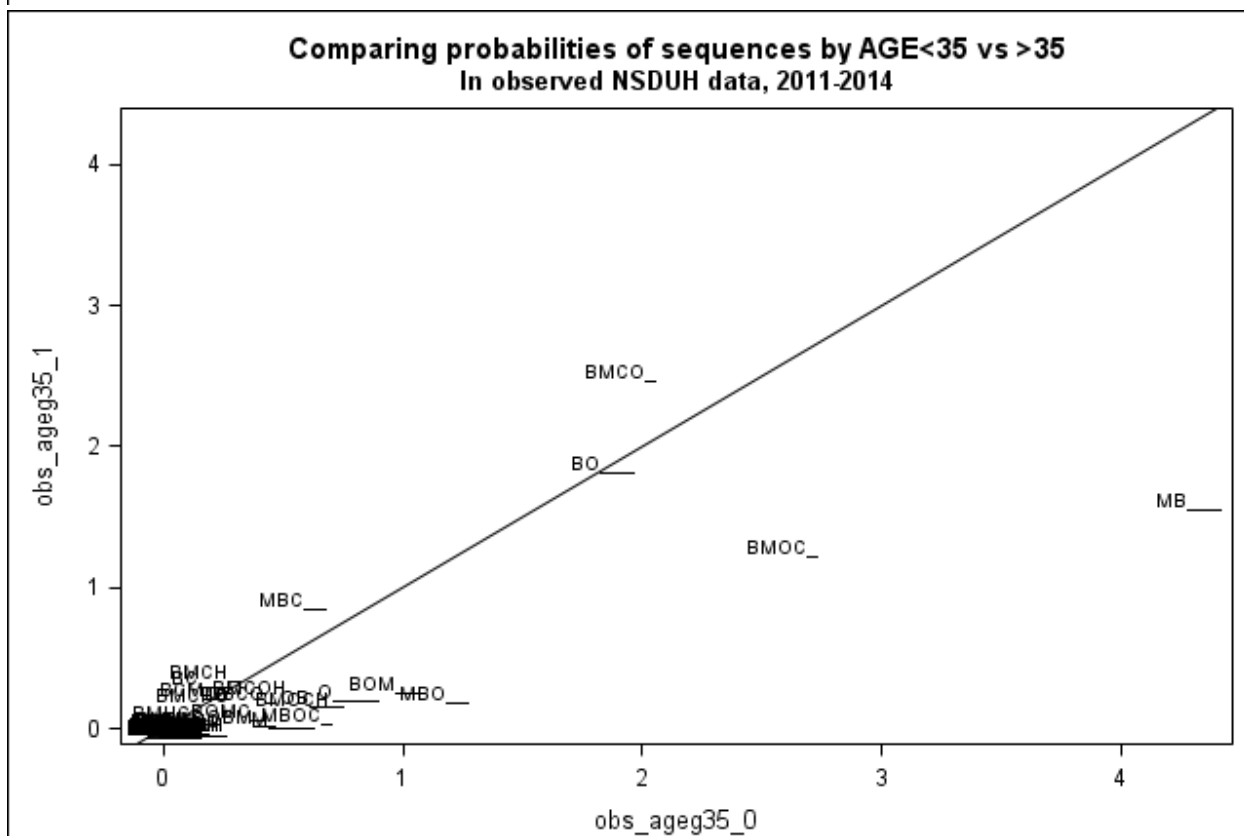
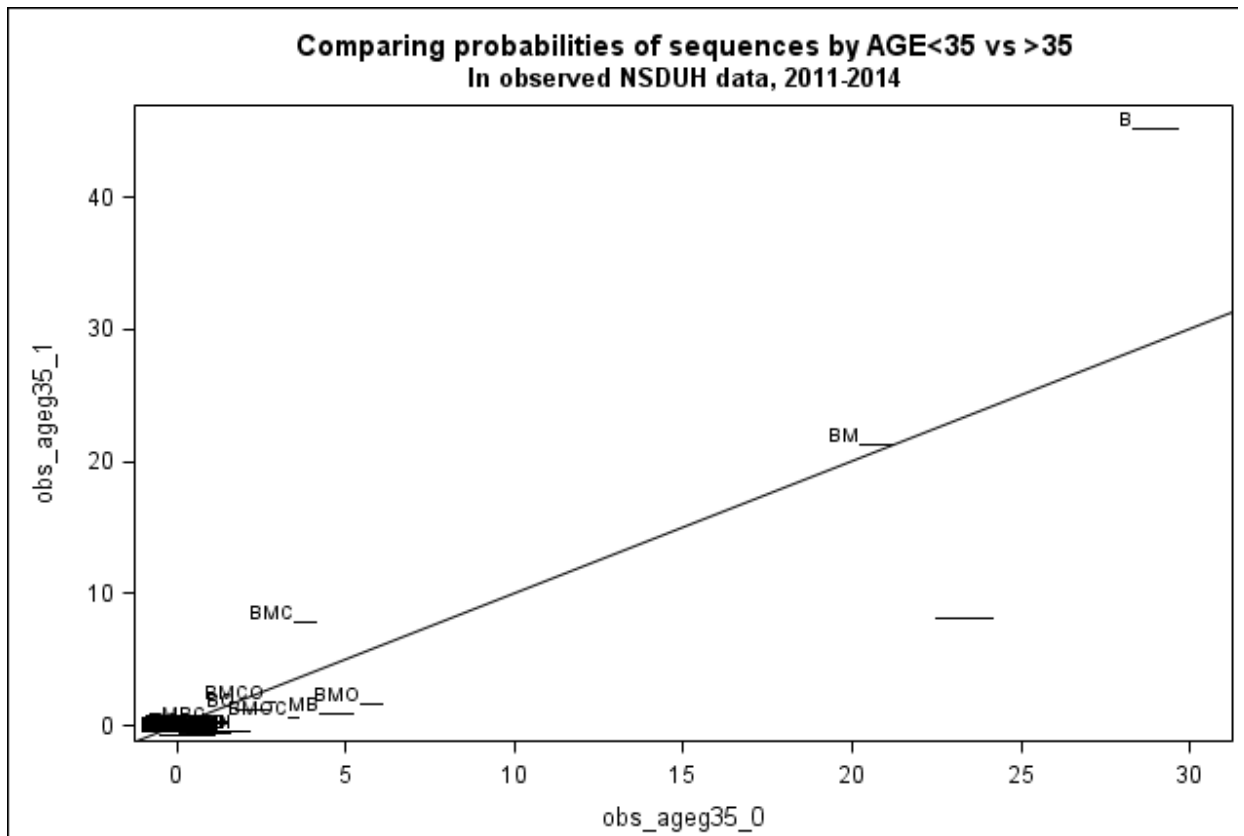
HMB_	0.0001	1
HOB_	0.0001	1
OCMB_	0.0001	1
OMHBC	0.0001	1
OBCM_H	0.0001	1
CHBM_	0.0001	1
HMBO_	0.000044	1
BOHCM	0.000036	1
MHOBC	0.000034	1
CHO_	0.000023	1
OHCM_	0.000022	1
CMHO_	0.000005	1

Sub group comparisons. The prevalence of each sequence was also calculated within strata of age and gender. The following graphs are summaries comparing the observed prevalence of sequences across these covariates within the observed data.

1. Gender- axes are percentages, bottom graph shows low end of the range.



2. Age at interview ≤ 34 vs. ≥ 35 . Axes show percentages, bottom graph shows low end of range.



Questions/issues:

- Method for 5-way ties “ok” given extreme rareness or need to do something more sophisticated?
Can be done but maybe not worth the time/effort.
- One imputation “ok”?
- Calculation of new “expected” sequence frequencies – we should just calculate one set of expected values each for the independent and correlated data sets, to be treated as constants for comparison with all observed (overall and subgroups) ... versus any subgroup-specific expected values....
 1. 2011-2014 only versus all years
- Best working list of subgroups that we want to look at?
 1. Male/female
 2. Age over/under 35 at survey
 3. Early vs. later age of onset – how should this be defined? Is there a ‘usual’ definition?
 4. Time trends- early/late (need earlier years of data)
 5. Cohort based on birth year (approx. born before or after 1980)