2020-02-25 15:28:25

I'll summarize the results again, after our initial chat. The main difference here is that we lost 4 subjects because I capped the participants to have at least 1 year between baseline and follow-up clinical assessments. I also have data files with subjects restricting it to 2 and 3 years between baseline and FU. Maybe we can run some robustness analysis with those later?

Here are the new descriptives:

```
> data0 =
readRDS('~/data/baseline_prediction/prs_start/complete_massagedResids_clin
DiffGE1_02202020.rds')
> table(data0$sex)

Female   Male
   128    261
> mean(data0$base_age)
[1] 8.206632
> sd(data0$base_age)
[1] 2.631743
> mean(data0$last_age)
[1] 13.09728
> sd(data0$last_age)
[1] 2.987265
> table(data0$thresh0.00_inatt_GE6_wp05)

     nv012        imp     nonimp notGE6adhd
       157        115         72         45
> table(data0$thresh0.50_hi_GE6_wp05)

     nv012        imp     nonimp notGE6adhd
       157         76        111         45
```

There are 6 data domains:

- neuropsych: 'FSIQ', "VMI.beery", "SSB.wisc", "SSF.wisc", 'DSF.wisc', 'DSB.wisc', "DS.wj", "VM.wj"
- demographic: 'base_age', 'sex', 'SES'
- genomics: PRS scores for the entire cohort (i.e. not the European-only PRS)
- DTI: FA values for JHU tracts, collapsed to reduce variables
- anatomy: thickness data for collapsed Freesurfer lobar regions
- clinics: 'internalizing', 'externalizing', 'base_inatt', 'base_hi'

Within each domain, prior to any sort of imputation, data were residualized. Here are the variables used for residualization within each domain:

- neuropsych: sex, base_age
- genomics: population PCs, base_age, sex

- DTI: "meanX.trans", "meanY.trans", "meanZ.trans", "meanX.rot", "meanY.rot", "meanZ.rot", "goodVolumes", age_at_scan, sex
- anatomy: "mprage_score", "ext_avg", "int_avg", age_at_scan, sex

Those variables were used initially, but the final model was optimized using stepAIC. Note that FSIQ, SES, externalizing, internalizing, and base_sx were NOT residualized.

Analysis 1: univariate results

Within each data domain, check which variables are significantly predicted in the model that uses the linear relationship between the 4 groups as the main predictor. In other words:

```
model = lme(myvar ~ ordered, ~1|FAMID)
```

And we collected the p-value and betas for the linear model associated with the variable **ordered**. The order of the groups is always ('nv012', 'notGE6adhd', 'imp', 'nonimp'). The number of observations varies per domain, as there was no imputation in this analysis. So, the final number per domain is as follows:

```
> idx=!is.na(data0[, 'FSIQ'])
> sum(idx)
[1] 386
> table(data0[idx,]$thresh0.00_inatt_GE6_wp05)

     nv012        imp     nonimp notGE6adhd
       155        114         72         45
> table(data0[idx,]$thresh0.50_hi_GE6_wp05)

     nv012        imp     nonimp notGE6adhd
       155         75        111         45
> idx=!is.na(data0[, 'VMI.beery'])
> sum(idx)
[1] 312
> table(data0[idx,]$thresh0.00_inatt_GE6_wp05)

     nv012        imp     nonimp notGE6adhd
       123         91         60         38
> table(data0[idx,]$thresh0.50_hi_GE6_wp05)

     nv012        imp     nonimp notGE6adhd
       123         60         91         38
> idx=!is.na(data0[, 'SSB.wisc'])
> sum(idx)
[1] 241
> table(data0[idx,]$thresh0.00_inatt_GE6_wp05)

     nv012        imp     nonimp notGE6adhd
        89         74         41         37
> table(data0[idx,]$thresh0.50_hi_GE6_wp05)
```

```
     nv012        imp     nonimp notGE6adhd
        89         49         66         37
> idx=!is.na(data0[, 'DS.wj'])
> sum(idx)
[1] 335
> table(data0[idx,]$thresh0.00_inatt_GE6_wp05)

     nv012        imp     nonimp notGE6adhd
       135        102         56         42
> table(data0[idx,]$thresh0.50_hi_GE6_wp05)

     nv012        imp     nonimp notGE6adhd
       135         66         92         42
> idx=!is.na(data0[, 'CC_fa'])
> sum(idx)
[1] 179
> table(data0[idx,]$thresh0.00_inatt_GE6_wp05)

     nv012        imp     nonimp notGE6adhd
        73         49         31         26
> table(data0[idx,]$thresh0.50_hi_GE6_wp05)

     nv012        imp     nonimp notGE6adhd
        73         33         47         26
> idx=!is.na(data0[, 'parietal'])
> sum(idx)
[1] 282
> table(data0[idx,]$thresh0.00_inatt_GE6_wp05)

     nv012        imp     nonimp notGE6adhd
       124         78         45         35
> table(data0[idx,]$thresh0.50_hi_GE6_wp05)

     nv012        imp     nonimp notGE6adhd
       124         53         70         35
```

Finally, the data used for anat and DTI are already the result of excluding outliers using the same methods we used in the heritability of change paper, after removing any scans outside the 95th percentile. Specifically, all longitudinal scans for the 389 subjects with PRS were taken into consideration. After removing the outliers, I picked the first scan and only kept it if it was acquired within 1 year of the baseline clinical assessment.

Results

So, the complete list of variables tested was:

```
ADHD_PRS (only the best PRS variable per SX was used in FDR)
frontal
parietal
cingulate
insula
temporal
```

```
occipital
OFC
sensorimotor
ATR_fa
CST_fa
CIN_fa
CC_fa
IFO_fa
ILF_fa
SLF_fa
UNC_fa
VMI.beery
SSB.wisc
SSF.wisc
DSF.wisc
DSB.wisc
DS.wj
VM.wj
FSIQ
SES
```

Also, the clinical domain variables, along with sex and age, were not taken into consideration for FDR. They'l make an appearance in the big model, but they skewed the results too much here to be considered in the univariayte analysis.

For multiple comparisons correction, our chosen approach was to do an FDR across all domains and SX. Then it's a matter of choosing what q level to go with. If I do Q < .1, which is acceptable as this is mostly for screening, we get:

```
> ps[p2<.1,c('brainVar', 'outcome')]
           brainVar                       outcome
1              FSIQ ORDthresh0.00_inatt_GE6_wp05
2         VMI.beery ORDthresh0.00_inatt_GE6_wp05
3             VM.wj ORDthresh0.00_inatt_GE6_wp05
4  ADHD_PRS0.000500 ORDthresh0.00_inatt_GE6_wp05
36        VMI.beery    ORDthresh0.50_hi_GE6_wp05
37             FSIQ    ORDthresh0.50_hi_GE6_wp05
38            VM.wj    ORDthresh0.50_hi_GE6_wp05
39           IFO_fa    ORDthresh0.50_hi_GE6_wp05
40           CST_fa    ORDthresh0.50_hi_GE6_wp05
41 ADHD_PRS0.001000    ORDthresh0.50_hi_GE6_wp05
42           ATR_fa    ORDthresh0.50_hi_GE6_wp05
43              OFC    ORDthresh0.50_hi_GE6_wp05
```

I also tried running the uinivariate analysis for the 3-group and 2-group comparisons, but we only got significant results for some of the PRS variables and the neuropsych ones. That could easily be due to the high loss of samples that incurs as we make fewer groups available.

I also ran the univariate analysis to check whether the WNH-only PRS worked better than the whole group PRS. Note that similar to the whole-group PRS, the WNH PRS was also residualized using the population PCs. To make the comparisons more fair, I trimmed the number of participants to only imclude the WNH ones, and compared whether the PRS or PRSeur did a better job in univariate analysis. There's barely any difference in the 4-group analysis. Nothing significant in the 3-class analysis, but the main result is in the 2-class analysis, where only PRS has two nominal hits, while PRSeur has none.

Analysis 2: big model

This analysis had the goal of checking how well we can model the groups by combining all the "good" univariate variables from analysis 1.

As we cannot deal with NAs here, we decided to impute the data using as the base the 179 kids who have both PRS and DTI.

The model is a multinomial logistic regression that ignores the family term. It is also not ordered, because it performed better (i.e. higher ROC AUC) than the ordered model:

```
group ~ good_vars + covars
```

The good_vars came from analysis 1:

```
hi_vars = c('VMI.beery', 'FSIQ', 'VM.wj', 'IFO_fa', 'CST_fa',
            'ADHD_PRS0.001000', 'ATR_fa', 'OFC')
inatt_vars = c('FSIQ', 'VMI.beery', 'VM.wj',
               'ADHD_PRS0.000500')
covars = c('base_age', 'sex')
```

As before, every domain has already been residualized within domain (for example, PRS was residualized based on PCs, etc). We can evaluate the models based on ROC AUC, and check how important each variable was in the prediction (just the sum of the absolute value of the coefficients across the different categories).

```
                  Overall
FSIQ             0.7202945
VMI.beery        0.7291959
VM.wj            1.2683058
ADHD_PRS0.000500 0.5230646
base_age         1.1259660
sexMale          1.6162441
Multi-class area under the curve: 0.6748

[1] "hi"
                  Overall
VMI.beery        0.6218178
FSIQ             0.7530228
VM.wj            1.1535166
IFO_fa           0.5700479
```

```
CST_fa           0.5163395
ADHD_PRS0.001000 0.4819661
ATR_fa           0.3725260
OFC              0.5520842
base_age         0.8095592
sexMale          1.5286176
Multi-class area under the curve: 0.6719
```

Results are not impressive. But since I have already removed age and sex within domain, how does the model perform if I don't use those covariates?

```
[1] "inatt"
                  Overall
FSIQ             0.7668251
VMI.beery        0.6310876
VM.wj            1.1767705
ADHD_PRS0.000500 0.4498268
Multi-class area under the curve: 0.6417

[1] "hi"
                  Overall
VMI.beery        0.5263482
FSIQ             0.7849392
VM.wj            1.1261404
IFO_fa           0.5583678
CST_fa           0.5051897
ADHD_PRS0.001000 0.4199229
ATR_fa           0.3926108
OFC              0.5459563
Multi-class area under the curve: 0.6571
```

We take a small hit, which somewhat makes sense as they can explain some variance in the outcomes that's not related to the domain-specific variables.

Of course, adding the clinical variables will offset everything in the 4-group comparison, but let's see how it goes anyways:

```
[1] "inatt"
                                      Overall
FSIQ                                41.982794
VMI.beery                            3.578685
VM.wj                               81.447472
ADHD_PRS0.000500                    43.894476
base_age                            75.004177
sexMale                            268.721519
externalizing1                     387.633135
internalizing1                     661.230632
medication_status_at_observationstim 266.648934
base_inatt                         763.886003
```

```
base_hi                                604.436176
Multi-class area under the curve: 0.9404

[1] "hi"
                                        Overall
VMI.beery                              60.808914
FSIQ                                    5.487375
VM.wj                                  14.009715
IFO_fa                                 51.110941
CST_fa                                  3.580805
ADHD_PRS0.001000                      102.269757
ATR_fa                                 67.143582
OFC                                    51.124880
base_age                              106.437983
sexMale                               194.399110
externalizing1                        153.551182
internalizing1                        815.631843
medication_status_at_observationstim 200.047641
base_inatt                            641.697137
base_hi                               605.877102
Multi-class area under the curve: 0.9485
```

Adding the clinical variables is almost unfair, given how the groups are defined on them. But it gives us an idea of how things play out in the 4-group case. Let's start removing groups and check the variable contributions.

```
[1] "inatt"
                   Overall
FSIQ             0.74308377
VMI.beery        0.50205321
VM.wj            0.09042903
ADHD_PRS0.000500 0.44625015
base_age         0.73738052
sexMale          0.31167702
Multi-class area under the curve: 0.6711

[1] "hi"
                  Overall
VMI.beery        0.4223979
FSIQ             0.7085519
VM.wj            0.1774528
IFO_fa           0.4794401
CST_fa           0.5468075
ADHD_PRS0.001000 0.4705216
ATR_fa           0.2503641
OFC              0.1618881
base_age         0.2792004
sexMale          0.4574787
Multi-class area under the curve: 0.6687
```

Results in 3 group analysis weren't that different than 4-group, without clinicals.

```
[1] "inatt"
                                     Overall
FSIQ                                0.8298191
VMI.beery                           0.1827814
VM.wj                               0.4307779
ADHD_PRS0.000500                    0.7593407
base_age                            0.5604268
sexMale                             0.3612244
externalizing1                      1.1293731
internalizing1                      1.2733140
medication_status_at_observationstim 1.1717559
base_inatt                          5.0131163
base_hi                             4.4278342
Multi-class area under the curve: 0.8808

[1] "hi"
                                     Overall
VMI.beery                           1.2465686
FSIQ                                1.3111484
VM.wj                               0.1714609
IFO_fa                              0.7547630
CST_fa                              0.5800093
ADHD_PRS0.001000                    0.7842605
ATR_fa                              1.3417393
OFC                                 0.6085251
base_age                            0.5490014
sexMale                             0.7308607
externalizing1                      3.8110658
internalizing1                      2.2393427
medication_status_at_observationstim 1.9101648
base_inatt                          5.5138883
base_hi                             6.5466645
Multi-class area under the curve: 0.8971
```

Results in 3-group classification are still quite decent. The variable contribution ratio is still quite disproportional though!

Let's play with the 2-group results:

```
[1] "inatt"
                  Overall
FSIQ            0.3448425
VMI.beery       0.1183606
VM.wj           0.1405499
ADHD_PRS0.000500 0.1915836
base_age        0.7591428
sexMale         0.2967667
Multi-class area under the curve: 0.7187

[1] "hi"
```

```
                     Overall
VMI.beery        0.23957522
FSIQ             0.09309914
VM.wj            0.14226124
IFO_fa           0.07215796
CST_fa           0.57775107
ADHD_PRS0.001000 0.23489728
ATR_fa           0.13699991
OFC              0.12243041
base_age         0.25246825
sexMale          0.34083300
Multi-class area under the curve: 0.7059
```

The 2-class results actually go a little over .7, which is nice to see.

```
[1] "inatt"
                                      Overall
FSIQ                                0.39183744
VMI.beery                           0.05199274
VM.wj                               0.43267531
ADHD_PRS0.000500                    0.19136429
base_age                            0.51344091
sexMale                             0.33444265
externalizing1                      0.15260005
internalizing1                      1.25503158
medication_status_at_observationstim 0.79584536
base_inatt                          1.20098822
base_hi                             0.01580077
Multi-class area under the curve: 0.8216

[1] "hi"
                                      Overall
VMI.beery                           0.68759308
FSIQ                                0.02644463
VM.wj                               0.16887468
IFO_fa                              0.15639033
CST_fa                              0.49915548
ADHD_PRS0.001000                    0.32403480
ATR_fa                              0.05469934
OFC                                 0.05350514
base_age                            0.51261043
sexMale                             0.70984814
externalizing1                      1.09603122
internalizing1                      2.11730148
medication_status_at_observationstim 0.68840197
base_inatt                          0.26609419
base_hi                             2.12790071
Multi-class area under the curve: 0.8391
```

So, using the clinical domain it wasn't so bad for the 2-class case. Yes, it's all training data, but there's something there.

Analysis 3: ML

The idea here is to take all variables that were analyzed in the univariate analysis (#1), but instead of taking them individually we take them all together (after any within-domain residualizing procedures).

There is no imputation because the classifiers are trained within domain. We separate for testing everyone but one participant in the same family. Some of the testing cases will have data only for some of the domains, similarly to what we will have in the training data. Note that the testing data is never used during training, but it's not a clean cross-validation: the test data is not independent from the training data because of the family component, and also because of the residualizing procedure that uses the entire dataset for robustness.

So, we train the best classifier we can within each domain. We also train an ensemble classifier that learns to combine the "vote" for each domain. In other words, each domain votes (with a probability) what group a given participant belongs to, and the ensemble classifier learns how to best consider each vote (i.e. trust/take into consideration some domains more than others). When there is no data for a given domain it either votes NA, or just the class probability deducted from the training data. I tried it both ways, the difference being that if voting NA we need to use an ensembler that takes that in (i.e. any GLM/weighted majority voting won't work).

The training itself is a 10-fold repeated cross validation (10 times), which happens only within the training set. For this analysis, we can not only assess variable importance within domain, but also how important each domain was in the ensemble classifier.

I only ran this for the 2 class scenarios, as I didn't think it'd be fair to run externalizing and medication variables in the 4 and 3-class cases.

For the 2-class case, we get ROC AUC up to .78/.7 for inatt and .63/.73 for hi, depending on how we ensemble the domains.

These are the training/testing splits in each domain (neuropsych was further divided to avoid additional imputation):

```
[1] "Training iq_vmi on thresh0.00_inatt_GE6_wp05 (sx=inatt,
model=kernelpls)"
[1] "Training on 103 participants"
[1] "Training wisc on thresh0.00_inatt_GE6_wp05 (sx=inatt,
model=kernelpls)"
[1] "Training on 72 participants"
[1] "Training wj on thresh0.00_inatt_GE6_wp05 (sx=inatt, model=kernelpls)"
[1] "Training on 106 participants"
[1] "Training demo on thresh0.00_inatt_GE6_wp05 (sx=inatt,
model=kernelpls)"
[1] "Training on 133 participants"
[1] "Training clin on thresh0.00_inatt_GE6_wp05 (sx=inatt,
model=kernelpls)"
[1] "Training on 133 participants"
[1] "Training gen on thresh0.00_inatt_GE6_wp05 (sx=inatt,
model=kernelpls)"
```

```
[1] "Training on 133 participants"
[1] "Training dti on thresh0.00_inatt_GE6_wp05 (sx=inatt,
model=kernelpls)"
[1] "Training on 56 participants"
[1] "Training anat on thresh0.00_inatt_GE6_wp05 (sx=inatt,
model=kernelpls)"
[1] "Training on 84 participants"
[1] "iq_vmi"
[1] "Testing on 48 participants"
[1] "wisc"
[1] "Testing on 42 participants"
[1] "wj"
[1] "Testing on 49 participants"
[1] "demo"
[1] "Testing on 54 participants"
[1] "clin"
[1] "Testing on 54 participants"
[1] "gen"
[1] "Testing on 54 participants"
[1] "dti"
[1] "Testing on 24 participants"
[1] "anat"
[1] "Testing on 39 participants"
```

```
[1] "Training iq_vmi on thresh0.50_hi_GE6_wp05 (sx=hi, model=kernelpls)"
[1] "Training on 103 participants"
[1] "Training wisc on thresh0.50_hi_GE6_wp05 (sx=hi, model=kernelpls)"
[1] "Training on 72 participants"
[1] "Training wj on thresh0.50_hi_GE6_wp05 (sx=hi, model=kernelpls)"
[1] "Training on 106 participants"
[1] "Training demo on thresh0.50_hi_GE6_wp05 (sx=hi, model=kernelpls)"
[1] "Training on 133 participants"
[1] "Training clin on thresh0.50_hi_GE6_wp05 (sx=hi, model=kernelpls)"
[1] "Training on 133 participants"
[1] "Training gen on thresh0.50_hi_GE6_wp05 (sx=hi, model=kernelpls)"
[1] "Training on 133 participants"
[1] "Training dti on thresh0.50_hi_GE6_wp05 (sx=hi, model=kernelpls)"
[1] "Training on 56 participants"
[1] "Training anat on thresh0.50_hi_GE6_wp05 (sx=hi, model=kernelpls)"
[1] "Training on 84 participants"
[1] "iq_vmi"
[1] "Testing on 48 participants"
[1] "wisc"
[1] "Testing on 42 participants"
[1] "wj"
[1] "Testing on 49 participants"
[1] "demo"
[1] "Testing on 54 participants"
[1] "clin"
[1] "Testing on 54 participants"
[1] "gen"
```

```
[1] "Testing on 54 participants"
[1] "dti"
[1] "Testing on 24 participants"
[1] "anat"
[1] "Testing on 39 participants"
```

Then it's a matter of choosing what sort of voting ensemlber we think makes more sense. First, let's look at variable importance for the case where we ignore the domain if the participant has no data, inattention firts:

```
[1] "iq_vmi"
[1] "Testing on 48 participants"
ROC curve variable importance

          Importance
VMI.beery        100
FSIQ               0
[1] "wisc"
[1] "Testing on 42 participants"
ROC curve variable importance

         Importance
SSB.wisc     100.00
DSB.wisc      34.82
DSF.wisc      10.53
SSF.wisc       0.00
[1] "wj"
[1] "Testing on 49 participants"
ROC curve variable importance

      Importance
DS.wj        100
VM.wj          0
[1] "demo"
[1] "Testing on 54 participants"
ROC curve variable importance

         Importance
base_age     100.00
SES           42.81
sex            0.00
[1] "clin"
[1] "Testing on 54 participants"
ROC curve variable importance

                                Importance
base_inatt                         100.000
base_hi                             45.063
medication_status_at_observation     4.754
externalizing                        1.788
internalizing                        0.000
[1] "gen"
```

```
[1] "Testing on 54 participants"
ROC curve variable importance

                Importance
ADHD_PRS0.000100    100.000
ADHD_PRS0.000050     88.644
ADHD_PRS0.000500     68.770
ADHD_PRS0.001000     39.905
ADHD_PRS0.050000     32.808
ADHD_PRS0.100000     14.196
ADHD_PRS0.005000     10.410
ADHD_PRS0.200000     10.095
ADHD_PRS0.010000      9.779
ADHD_PRS0.400000      3.943
ADHD_PRS0.500000      1.262
ADHD_PRS0.300000      0.000
[1] "dti"
[1] "Testing on 24 participants"
ROC curve variable importance

       Importance
CIN_fa    100.000
CST_fa     96.053
UNC_fa     46.053
CC_fa      43.421
SLF_fa     38.158
IFO_fa     31.579
ILF_fa      2.632
ATR_fa      0.000
[1] "anat"
[1] "Testing on 39 participants"
ROC curve variable importance

             Importance
frontal         100.000
insula           65.942
OFC              64.493
temporal         58.696
cingulate        12.319
parietal          9.420
occipital         3.623
sensorimotor      0.000
      ROC      Sens      Spec
0.7000000 0.5000000 0.7857143
C5.0Tree variable importance

       Overall
clin    100.00
gen      72.93
wj       47.37
anat     12.78
dti       0.00
demo      0.00
wisc      0.00
```

```
iq_vmi    0.00
[1] "inatt,hdda,C5.0Tree,0.919770,0.700000"
```

So, our testing ROC AUC is .70. Now missing votes for hi:

```
[1] "iq_vmi"
[1] "Testing on 48 participants"
ROC curve variable importance

          Importance
FSIQ             100
VMI.beery          0
[1] "wisc"
[1] "Testing on 42 participants"
ROC curve variable importance

         Importance
SSF.wisc    100.000
DSB.wisc     12.613
DSF.wisc      1.802
SSB.wisc      0.000
[1] "wj"
[1] "Testing on 49 participants"
ROC curve variable importance

      Importance
VM.wj        100
DS.wj          0
[1] "demo"
[1] "Testing on 54 participants"
ROC curve variable importance

         Importance
base_age     100.00
sex           91.13
SES            0.00
[1] "clin"
[1] "Testing on 54 participants"
ROC curve variable importance

                                Importance
base_hi                        100.0000
base_inatt                       8.8206
internalizing                    1.4113
medication_status_at_observation     0.7056
externalizing                    0.0000
[1] "gen"
[1] "Testing on 54 participants"
ROC curve variable importance

                 Importance
ADHD_PRS0.000100    100.000
```

```
ADHD_PRS0.000500     69.553
ADHD_PRS0.001000     68.715
ADHD_PRS0.010000     65.642
ADHD_PRS0.000050     63.128
ADHD_PRS0.005000     60.335
ADHD_PRS0.300000     18.994
ADHD_PRS0.400000      7.542
ADHD_PRS0.050000      7.263
ADHD_PRS0.500000      4.190
ADHD_PRS0.200000      2.793
ADHD_PRS0.100000      0.000
[1] "dti"
[1] "Testing on 24 participants"
ROC curve variable importance

       Importance
CST_fa    100.000
UNC_fa     57.297
ILF_fa     51.892
CC_fa      48.649
IFO_fa     20.000
SLF_fa     14.054
CIN_fa      1.081
ATR_fa      0.000
[1] "anat"
[1] "Testing on 39 participants"
ROC curve variable importance

             Importance
OFC              100.00
parietal          51.74
cingulate         47.76
frontal           26.87
insula            16.92
sensorimotor      16.42
temporal          10.45
occipital          0.00
      ROC      Sens      Spec
0.7270233 0.4074074 0.9259259
C5.0Tree variable importance

       Overall
clin    100.00
dti      26.32
anat     23.31
demo      0.00
wj        0.00
gen       0.00
wisc      0.00
iq_vmi    0.00
[1] "hi,hdda,C5.0Tree,0.846088,0.727023"
```

Our testing ROC AUC is .73.

Now we try domain voting imputation based on class train ratios. Inatt first:

```
[1] "iq_vmi"
ROC curve variable importance

          Importance
VMI.beery        100
FSIQ               0
[1] "wisc"
ROC curve variable importance

         Importance
SSB.wisc     100.00
DSB.wisc      34.82
DSF.wisc      10.53
SSF.wisc       0.00
[1] "wj"
ROC curve variable importance

      Importance
DS.wj        100
VM.wj          0
[1] "demo"
ROC curve variable importance

         Importance
base_age     100.00
SES           42.81
sex            0.00
[1] "clin"
ROC curve variable importance

                                Importance
base_inatt                         100.000
base_hi                             45.063
medication_status_at_observation     4.754
externalizing                        1.788
internalizing                        0.000
[1] "gen"
ROC curve variable importance

                 Importance
ADHD_PRS0.000100    100.000
ADHD_PRS0.000050     88.644
ADHD_PRS0.000500     68.770
ADHD_PRS0.001000     39.905
ADHD_PRS0.050000     32.808
ADHD_PRS0.100000     14.196
ADHD_PRS0.005000     10.410
ADHD_PRS0.200000     10.095
ADHD_PRS0.010000      9.779
ADHD_PRS0.400000      3.943
ADHD_PRS0.500000      1.262
```

```
ADHD_PRS0.300000      0.000
[1] "dti"
ROC curve variable importance

       Importance
CIN_fa    100.000
CST_fa     96.053
UNC_fa     46.053
CC_fa      43.421
SLF_fa     38.158
IFO_fa     31.579
ILF_fa      2.632
ATR_fa      0.000
[1] "anat"
ROC curve variable importance

             Importance
frontal         100.000
insula           65.942
OFC              64.493
temporal         58.696
cingulate        12.319
parietal          9.420
occipital         3.623
sensorimotor      0.000
rpart2 variable importance

       Overall
clin   100.000
gen     62.971
demo    49.842
iq_vmi  29.977
dti     29.127
wisc    16.268
anat     4.765
wj       0.000
[1] "inatt,hdda,rpart2,0.930345,0.782143"
```

It's the highest AUC ROC we get, at .78, and the ensemble looks a bit more equalitarian too.

And also voting imputations for hi:

```
[1] "iq_vmi"
ROC curve variable importance

          Importance
FSIQ             100
VMI.beery          0
[1] "wisc"
ROC curve variable importance

         Importance
```

```
SSF.wisc    100.000
DSB.wisc     12.613
DSF.wisc      1.802
SSB.wisc      0.000
[1] "wj"
ROC curve variable importance

      Importance
VM.wj        100
DS.wj          0
[1] "demo"
ROC curve variable importance

         Importance
base_age     100.00
sex           91.13
SES            0.00
[1] "clin"
ROC curve variable importance

                                Importance
base_hi                           100.0000
base_inatt                          8.8206
internalizing                       1.4113
medication_status_at_observation    0.7056
externalizing                       0.0000
[1] "gen"
ROC curve variable importance

                 Importance
ADHD_PRS0.000100    100.000
ADHD_PRS0.000500     69.553
ADHD_PRS0.001000     68.715
ADHD_PRS0.010000     65.642
ADHD_PRS0.000050     63.128
ADHD_PRS0.005000     60.335
ADHD_PRS0.300000     18.994
ADHD_PRS0.400000      7.542
ADHD_PRS0.050000      7.263
ADHD_PRS0.500000      4.190
ADHD_PRS0.200000      2.793
ADHD_PRS0.100000      0.000
[1] "dti"
ROC curve variable importance

       Importance
CST_fa    100.000
UNC_fa     57.297
ILF_fa     51.892
CC_fa      48.649
IFO_fa     20.000
SLF_fa     14.054
CIN_fa      1.081
ATR_fa      0.000
```

```
[1] "anat"
ROC curve variable importance

             Importance
OFC              100.00
parietal          51.74
cingulate         47.76
frontal           26.87
insula            16.92
sensorimotor      16.42
temporal          10.45
occipital          0.00
rpart2 variable importance

       Overall
clin    100.00
dti      75.65
demo     66.89
wisc     58.73
anat     35.93
gen      27.17
iq_vmi   14.53
wj        0.00
[1] "hi,hdda,rpart2,0.814626,0.631001"
```

Results not as strong, though.

TODO

- robustness analysis using subjects with 2 and 3 years between baseline and follow-up?
- voxelwise analysis for the brain?