Leveraging multimodal data to predict outcomes of antidepressant treatment

Statistical analysis report

Brice Ozenne

April 22, 2022

1 Data

Data: has been provided by Emily Beaman. She processed relevant data from the CIMBI database leading to n=98 patients. In addition we computed:

- a "MR" biomarker as the average of the thickness of the left lateral, right lateral, left medial, and right medial orbitol frontal cortex.
- a "PET" biomarker by combining the log-binding potential from neocortex, hippocampus, caudate, and putamen via a latent variable model (adjusted for age, gender, and injected mass). This was done in a leave-one-out fashion i.e. the biomarker for individual i was obtained by fitting the model on all but the i-th individual and estimating the latent variable for the i-th individual based on its log-binding in the 4 regions.
- a "cognition" biomarker was obtained via a k-means algorithm on various cognitive outcomes (no leave-one-out here).

Outcome: relative change in HAMD6 between week 4/8/12 and baseline smaller or equal to 50% (recovery vs. no recovery).

Missing outcome data: among the 98 patients:

- 11 had a missing outcome at week 4. Patient 56123 had a spontaneous remission and was classified as "recovery" at week 4/8/12. Patient 55981 was hopsitalized (suicidal and psychotic) and was classified as "no recovery" at week 4/8/12.
- 14 had a missing outcome at week 8. Patients 56123 and 55981 were handled in the same way as week 4. Patient 55815 expericenced side effects to SNRI and 55851 attempt suicide. Both were classified as "no recovery" at week 8/12.

- 17 had a missing outcome at week 12. Patients 56123, 55981, 55815, and 55851 were handled in the same way as week 8. Patient 55742 had psychotic depression and was classified as "no recovery" at week 12.

8 patients with missing outcomes at all weeks were excluded (1 patient with missing outcome at week 4 had an outcome at week 8). Other patients for which the outcome could not be determined were excluded from week-specific analyses.

The corresponding R code is in the file `0-data-management.R` available on Github.

2 Data analysis

After discussion with neuroscientists, we have identified 10 candidate biomarkers[1] for predicting recovery after SSRI treatment:

- `MR_OFCthick`: thickness of the OFC brain region measured with MR.
- `HAMD17`: depression score at baseline.
- `low_hsCRP`: high sensitivity CRP (l levels of inflammation in the body).
- `lvpet`: summary of the brain log-PET binding.
- `cognitive_cluster`: summary of the cognition.
- `EEG_vigilance`: EEG signal (vigilance slope B1 bl).
- `CATS_scoretotal`: Childhood trauma (self reported).
- `CAR_AUCi`: Difference between the cortisol value and the cortisol value at wake-up cumulated over an hour.
- `neuroticism`: personality trait indexing tendency to experience negative emotion.

Missing biomarker values: two analyses were performed

- an analysis handling missing data either using multiple imputation (MI) in the association study or only using only the biomarkers with no missing value for the specific individual when computing predictions in the prediction study.
  When using MI, the original data was cloned 100 times. In each clone, missing values were imputed using Fully Conditional Specification (FCS) implemented by the MICE algorithm (Van Buuren and Groothuis-Oudshoorn, 2011). This algorithm alternates between learning the relationship between variables, using a linear regression for continuous variables, logistic regression for binary

[1] fMRI is missing in the list

variables, and a a proportional odds model for categorical variables with all variables (outcome, age, gender, biomarkers) as predictors (as suggested in Moons et al. (2006)), and impute by sampling from the resulting distributions (rougthly speaking, a noisy version of the best prediction). Pooling of the results was performed according Rubin's rule when computing the variance-covariance of the coefficients.
This was the primary analysis.

- a complete case analyses on 7 biomarkers (all but CATS, Cortisol, Neuroticism). This was the secondary/sensitivity analysis.

Association between recovery and biomarkers[2]: to assess whether the biomarker were associated with recovery we fitted a logistic regression with gender and age vs. a logistic regression with gender and age plus all biomarkers (as additive effects). Wald tests were used to test the association for each biomarker. P-value were adjusted for multiple comparison (i.e. FWER control) using a max-test adjustment. Random Forests (default hyperparameters: mtry = 3, 500 trees, node size 1) were also used to identify possible non-linear effects: a permutation test based on variable importance was performed.
This procedure was performed at week 4, week 8 and week 12. Post-hoc to the complete case results at week 12, A Generalized Additive Model (GAM) with age and gender as a linear effect and PET as non-linear was also performed at week 8 and 12.
For graphical display, the various logistic model were re-estimated after centering and scaling each variable and the corresponding estimates were displayed using a forest plot.

Predictive value of the biomarkers[3]: to assess whether the biomarkers can be used to discriminate between patients who will recover and patients who won't, we tested whether the Area Under the Curve (AUC) of a logistic model with age, gender, and the biomarkers was greated than 0.5, and whether it was greater than a model with only age and gender. We also compared the performance with a random forest (default hyperparameters: mtry = 3, 500 trees, node size 1). The AUC was estimated via 25 fold cross-validation repeted 100 times. An AUC was computed over all folds of a given repetition and then averaged acrossed repetions. The p-value was computed via a permutation test (10000 repetitions complete case, 1000 repetition missing values).

The corresponding R code is in a directory available on Github (files `analysis-XXX.R`).

[2] how does the recovery vary in average (i.e. at a population level) as a function of the biomarkers
[3] are the biomarkers useful to predict recovery for an individual

3 Results

3.1 Descriptive statistics

The dataset contained 90 patients, 89 with the outcome at week 4, 8 with outcome at week 8, and 86 with the outcome at week 12. Some summary statistics are displayed below:

```
    sex          age         MR_OFCthick        HAMD17       hsCRP        lvpet
male  :25   Min.   :18.24   Min.   :2.318   Min.   :18.00   high:19   Min.   :-0.82582
female:65   1st Qu.:22.11   1st Qu.:2.510   1st Qu.:20.00   low :69   1st Qu.:-0.48794
            Median :23.99   Median :2.566   Median :22.00   NA's: 2   Median :-0.42200
            Mean   :26.98   Mean   :2.576   Mean   :22.86             Mean   :-0.43020
            3rd Qu.:28.43   3rd Qu.:2.639   3rd Qu.:25.00             3rd Qu.:-0.35131
            Max.   :57.31   Max.   :2.889   Max.   :31.00             Max.   :-0.09773
                                                                      NA's   :2
cognitive_cluster EEG_vigilance      CATS_scoretotal    CAR_AUCi        neuroticism
Min.   :1.000     Min.   :-1.50000   Min.   : 0.0    Min.   :-1070.3   Min.   : 67.0
1st Qu.:1.000     1st Qu.: 0.00000   1st Qu.:16.0    1st Qu.:   79.1   1st Qu.:108.8
Median :2.000     Median : 0.00000   Median :23.0    Median :  221.9   Median :119.0
Mean   :1.875     Mean   :-0.01744   Mean   :30.1    Mean   :  181.3   Mean   :120.4
3rd Qu.:3.000     3rd Qu.: 0.00000   3rd Qu.:41.5    3rd Qu.:  381.1   3rd Qu.:134.0
Max.   :3.000     Max.   : 1.50000   Max.   :81.0    Max.   :  768.9   Max.   :155.0
NA's   :2         NA's   :4          NA's   :12      NA's   :21        NA's   :26
   Y_w4            Y_w8           Y_w12
Mode :logical   Mode :logical   Mode :logical
FALSE:52        FALSE:40        FALSE:26
TRUE :37        TRUE :48        TRUE :60
NA's :1         NA's :2         NA's :4
```

One biomarker looks a bit weird: `EEG_vigilance` with many 0 values:

```
table(dfWR.NP1$EEG_vigilance)
```

```
-1.5    -1 -0.75  -0.5 -0.25     0  0.25   0.5     1   1.5
   1     5     1     8     3    57     2     2     3     4
```

Should it be categorized: negative, null, positive?

The dataset contained many missing values. The pattern of the missing values is summarized on Figure 1. 50 patients had full data and the rest of the patients had between 1 and 5 missing data (number of red boxes per line). CATS, CAR, and neuroticm had a large number of missing data (12, 21, and 26) and this is why they were excluded from some analyses.

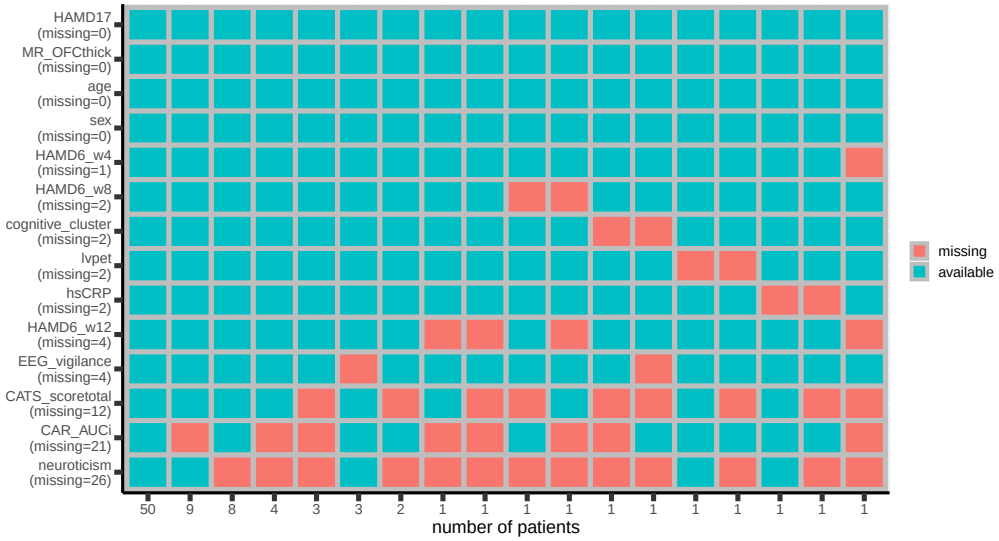

Figure 1: Missing data patterns

3.2 Outcome trajectories

The following table describes:

- at week 4: the number of patients that recovered (`nr2r`) or who did not recovered (`nr2nr`).
- at week 8 and 12: the number of patients who did not recover before or at the current time (`nr2nr`), the number of patients who just recovered (`nr2r`), the number of patients who recovered previously but go worse (`r2nr`), and the number of patients recovered previously and stay recovered (`r2r`). `nr2r+r2r` is then number of patients currently classified as recovered and `nr2nr+r2nr` as not recovered.

	week4	week8	week12
nr2nr	52 (58.43%)	30 (34.48%)	20 (23.53%)
r2nr	0 (0%)	9 (10.34%)	6 (7.06%)
nr2r	37 (41.57%)	20 (22.99%)	17 (20%)
r2r	0 (0%)	28 (32.18%)	42 (49.41%)
total	89 (100%)	87 (100%)	85 (100%)

To further describe the outcome trajectory of the patients over time, we use a latent class linear mixed model with homogeneous residual variance-covariance between groups to identify 3 groups of recovery. The results are shown in figure Figure 2 and ressemble those of Goerigk et al. (2021).

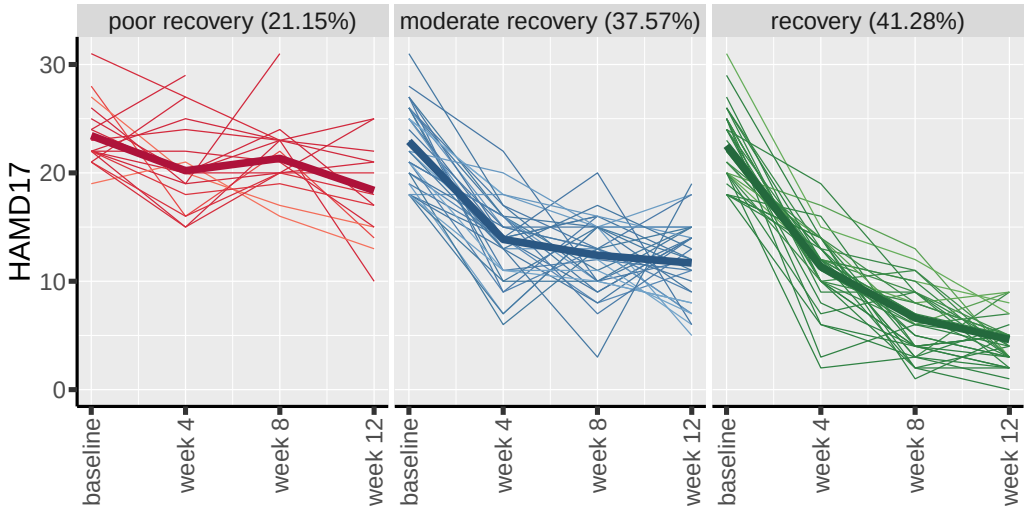

Figure 2: Recovery groups found by a latent class linear mixed model (LCMM). Thin lines represent individual trajectories colored as a function of the group membership probability. Thick lines represent group trajctories estimated by the LCMM.

3.3 Association study (linear)

The following table shows the result of multiple imputation for the logistic model with biomarkers at week 8, based on 88 patients (some with missing biomarker values):

	term	estimate	p.value	adj.p.value	lower.adj	upper.adj
1:	MR (OFC thickness)	-0.48611743	0.09664484	0.59435171	-1.3153682	0.3431334
2:	HAMD17	0.32097935	0.23911230	0.91059402	-0.4554627	1.0974214
3:	hsCRP	0.89399954	0.17305117	0.81271604	-0.9710296	2.7590287
4:	PET (serotonin)	-0.09616763	0.71280103	0.99999133	-0.8433506	0.6510154
5:	cognition (cluster 2)	-0.69912372	0.27221893	0.93996471	-2.5133247	1.1150773
6:	cognition (cluster 3)	-1.87992179	0.01004351	0.09046179	-3.9211639	0.1613203
7:	EEG (vigilance)	-0.69847488	0.02015285	0.17155060	-1.5426301	0.1456804
8:	CATS	0.07479048	0.78655856	0.99999954	-0.7153047	0.8648857
9:	Cortisol	0.28152530	0.34520600	0.97701174	-0.5691779	1.1322285
10:	Neuroticism	0.22612552	0.50133395	0.99819203	-0.7346358	1.1868868
11:	female	-0.57753491	0.33310871	NA	NA	NA
12:	age	0.45193136	0.16838786	NA	NA	NA

The smallest adjusted p-value is 0.09 obtained for cognition cluster 3: being is this cluser is associated with lower remission rate (estimate OR=0.15). The second most significant p-value is obtained for EEG.

Sensitivity analysis: we replicated this analysis at week 4:

	term	estimate	p.value	adj.p.value	lower.adj	upper.adj
1:	MR (OFC thickness)	-0.05876414	0.81995462	0.9999999	-0.7977520	0.6802237
2:	HAMD17	0.08812129	0.72970162	0.9999958	-0.6416895	0.8179321
3:	hsCRP	0.65041183	0.32292900	0.9707452	-1.2267548	2.5275784
4:	PET (serotonin)	0.08061616	0.74715773	0.9999978	-0.6349798	0.7962121
5:	cognition (cluster 2)	-0.33296344	0.58099358	0.9996948	-2.0579792	1.3920523
6:	cognition (cluster 3)	-1.30528058	0.05631125	0.4105926	-3.2382481	0.6276869
7:	EEG (vigilance)	-0.10138066	0.68695947	0.9999823	-0.8211744	0.6184131
8:	CATS	-0.28135338	0.31830848	0.9688761	-1.0857282	0.5230215
9:	Cortisol	0.28873615	0.34775562	0.9793286	-0.5888975	1.1663698
10:	Neuroticism	0.20053395	0.52825021	0.9990415	-0.7084860	1.1095539
11:	female	-1.15006496	0.04969227	NA	NA	NA
12:	age	0.24249711	0.35893909	NA	NA	NA

and week 12:

	term	estimate	p.value	adj.p.value	lower.adj	upper.adj
1:	MR (OFC thickness)	-0.87021372	0.01059592	0.09403706	-1.8211883	0.08076082
2:	HAMD17	0.22864772	0.43895218	0.99386231	-0.6141215	1.07141690
3:	hsCRP	0.54170359	0.48043219	0.99698507	-1.6490589	2.73246604
4:	PET (serotonin)	0.14969225	0.61394466	0.99981907	-0.6978764	0.99726089
5:	cognition (cluster 2)	-1.23228367	0.11663155	0.65755046	-3.4576855	0.99311813
6:	cognition (cluster 3)	-1.23211761	0.14918403	0.75139652	-3.6559188	1.19168357
7:	EEG (vigilance)	-0.05548535	0.85470244	0.99999999	-0.9215959	0.81062525
8:	CATS	0.69398065	0.04752293	0.34992677	-0.2933312	1.68129249
9:	Cortisol	0.46282495	0.22519432	0.88903756	-0.6224248	1.54807472
10:	Neuroticism	-0.02338173	0.95340882	1.00000000	-1.1674326	1.12066918
11:	female	-0.91743988	0.18900657	NA	NA	NA
12:	age	0.73697667	0.13715573	NA	NA	NA

At week 4 we see again some evidence for cluster 3 (without adjustment for multiple comparisons). At week 12, the most promising biomarkers would be OFC thickness and CATS. We also replicated the analysis when removing CATS, Cortisol, Neuroticm and using a complete case approach. The results are summarized on the following forest plot (Figure 3). We note that for some variables (age, cognition cluster 2, OFC thickness) the association seemed more and more pronounced over the weeks.

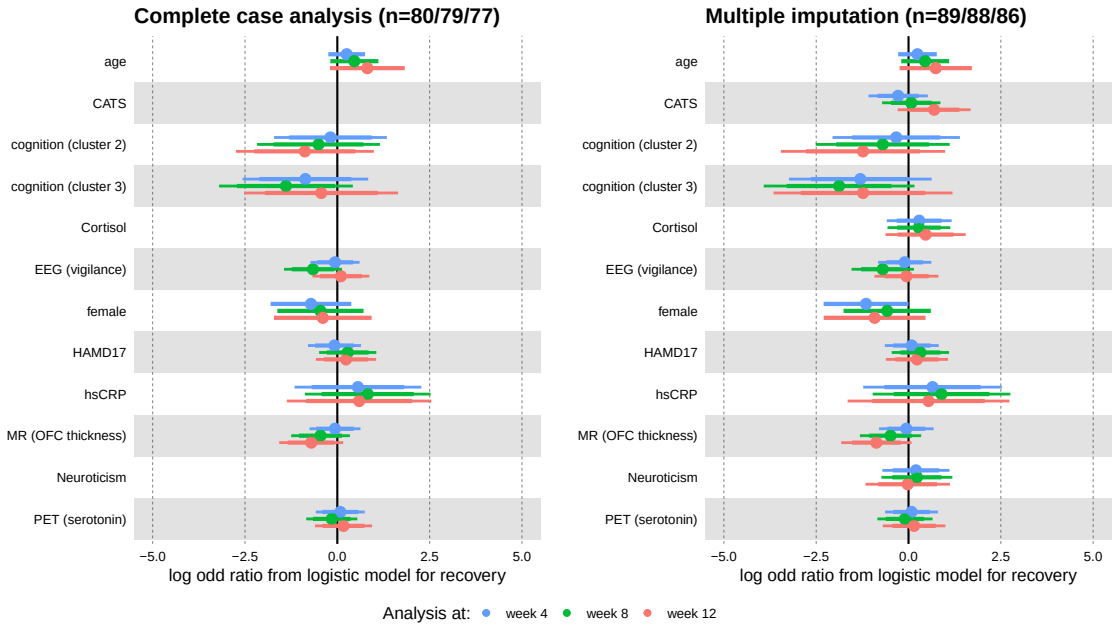

Figure 3: Log-odd ratio estimates (full circles), confidence intervals (thick lines) and adjusted confidence intervals (thin lines) for each analysis at each timepoint. Covariates have been centered and scale to be comparable. Adjustment for multiplicity is performed over biomarkers but not over time.

3.4 Association study (non-linear)

When using random forest with the complete case approach, no biomarker appeared relevant at week 8 (see also Figure 4):

	param	week4	week8	week12
1	female	-5e-04 (p=0.4535)	-0.0047 (p=0.8501)	-0.0025 (p=0.6713)
2	age	0.0112 (p=0.1698)	0.0038 (p=0.3257)	0.004 (p=0.3407)
3	MR (OFC thickness)	0.0029 (p=0.3586)	0.0084 (p=0.2328)	0.0134 (p=0.1319)
4	HAMD17	-0.0037 (p=0.6054)	0.0025 (p=0.3646)	4e-04 (p=0.4585)
5	hsCRP	-0.003 (p=0.7393)	-0.0011 (p=0.5255)	-0.0029 (p=0.7702)
6	PET (serotonin)	0.0075 (p=0.2438)	-0.0073 (p=0.6933)	0.0244 (p=0.026)
7	cognition (cluster 2)	-7e-04 (p=0.4845)	5e-04 (p=0.3696)	0.0043 (p=0.1518)
8	cognition (cluster 3)	0.0093 (p=0.0629)	0.0058 (p=0.1239)	-0.0026 (p=0.6643)
9	EEG (vigilance)	-0.0019 (p=0.5465)	0.0034 (p=0.2527)	0.0017 (p=0.2847)

At week 4, cognition (cluster) 3 was borderline significant. At week 12, there was some evidence for PET to be associated with the recovery.

Further investigation using splines reveal an inverted U-shape for the PET association (Figure 6, p=0.0128). A similar but non significant association was observed for PET at week 8 (Figure 5, p=0.0904).

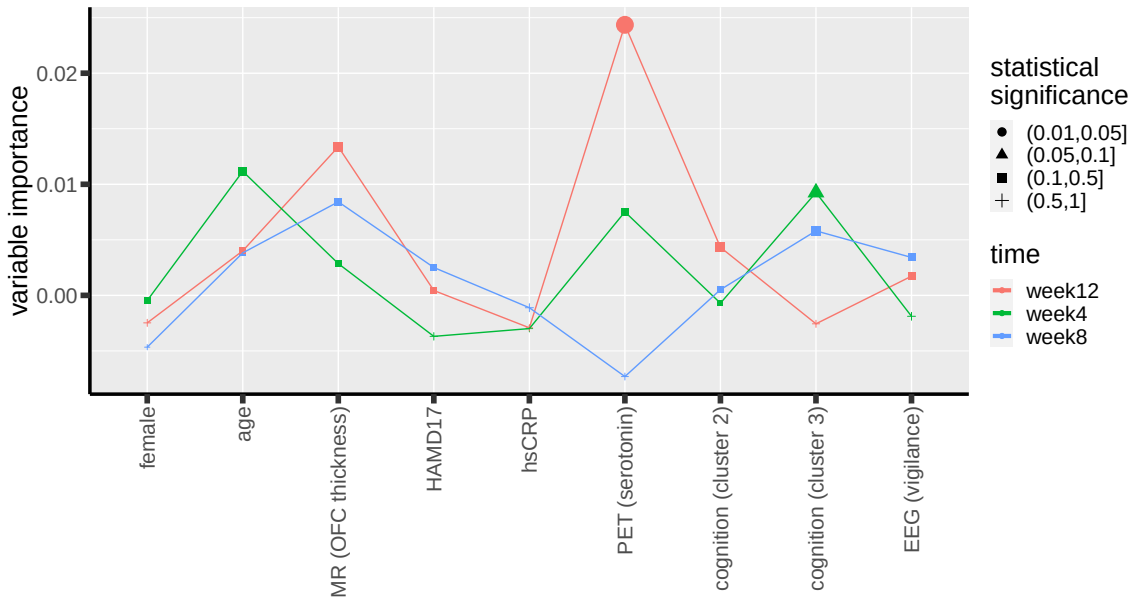

Figure 4: Log-odd ratio estimates (full circles), confidence intervals (thick lines) and adjusted confidence intervals (thin lines) for each analysis at each timepoint. Adjustment for multiplicity is performed over biomarkers but not over time.

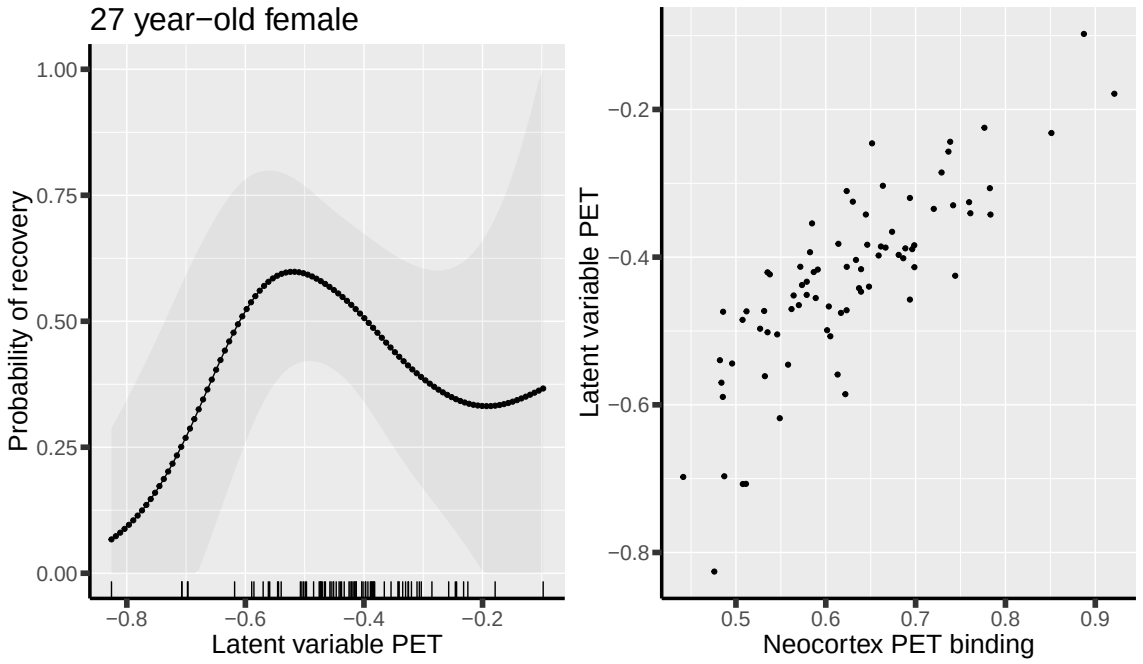

Figure 5: Association between recovery at week 8 and PET in a logistic model adjusted for age and sex (complete case).

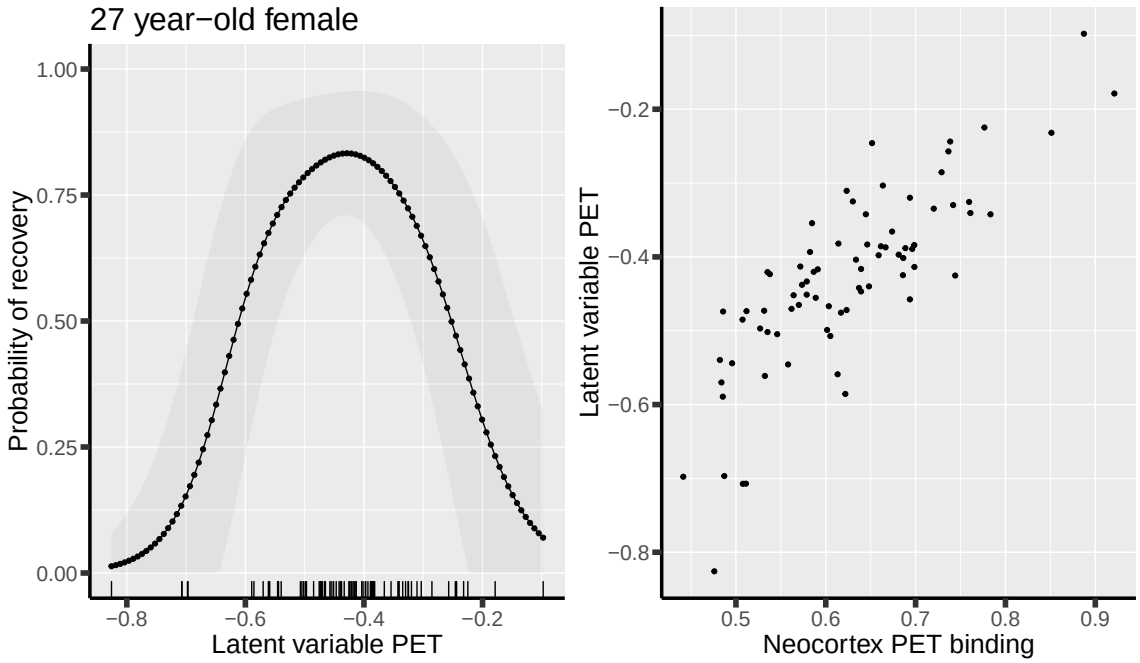

Figure 6: Association between recovery at week 12 and PET in a logistic model adjusted for age and sex (complete case).

3.5 Predictive value

Figure 7, Figure 8, and Figure 9 display the predicted probability obtain after cross-validation colored by recovery group. Overall it looks that the group are comparable at week 4 but there may be a difference at week 8 and 12. The corresponding ROC curve are put in appendix (Figure A, Figure B, Figure C).

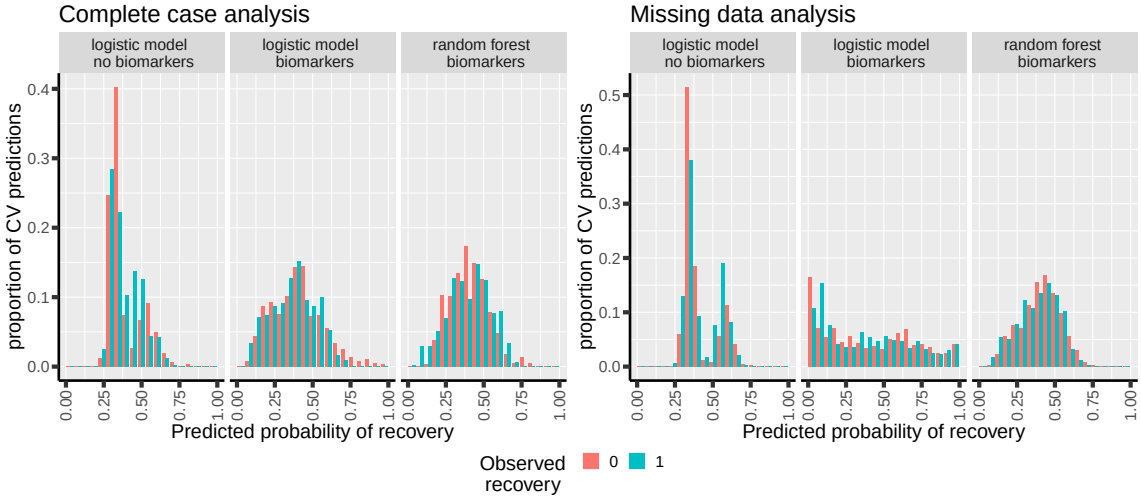

Figure 7: Distribution of the predicted probability of recovery at week 4 per group according to each model.

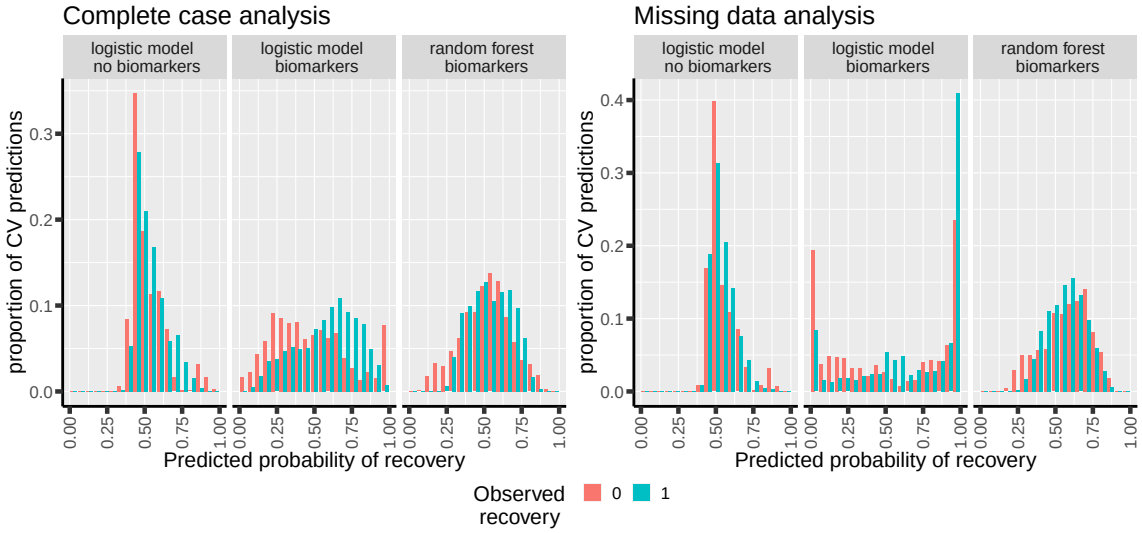

Figure 8: Distribution of the predicted probability of recovery at week 8 per group according to each model.

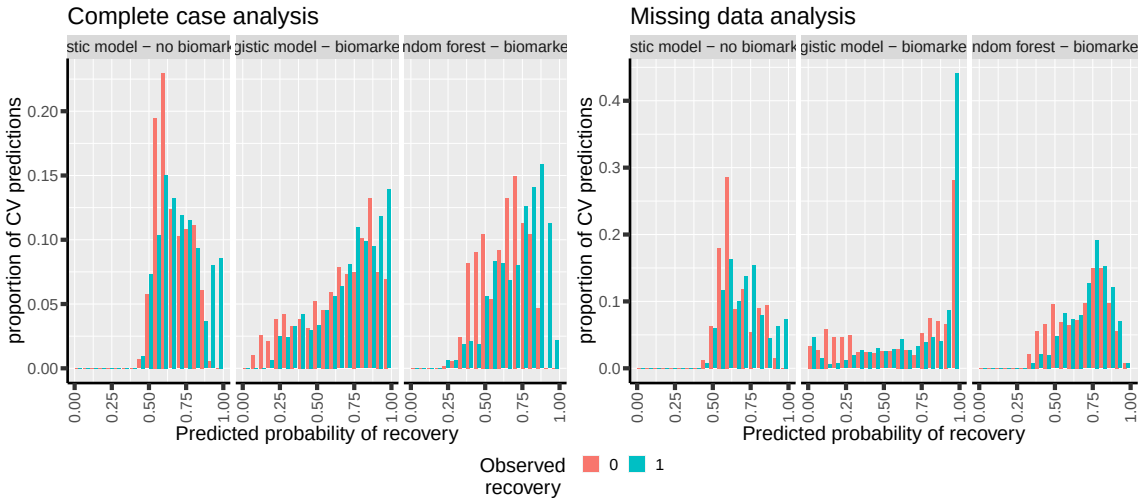

Figure 9: Distribution of the predicted probability of recovery at week 12 per group according to each model.

The table below summarizes the predictive performance. There was some evidence for some discriminative power of the biomarkers at week 8 (AUC=0.643, p=0.011) and 12 (AUC=0.573, p=0.119) with the logistic regression. The calibration of these model was however not clearly better than the null model with interept only (p=0.158 and p=0.063).

The point estimates of the random forest were not superior to the logistic model (with biomarkers) in term of AUC. The brier score was comparable to the logistic model with no biomarker.

```
                model week             AUC           Brier
1: GLM (no biomarker)    4 0.488 (p=0.226) 0.249 (p=0.161)
2:                       8 0.513 (p=0.162) 0.259 (p=0.352)
3:                      12 0.608 (p=0.018) 0.207 (p=0.026)
4:   GLM (biomarkers)    4 0.489 (p=0.498) 0.339 (p=0.788)
5:                       8 0.643 (p=0.011) 0.311 (p=0.158)
6:                      12 0.605 (p=0.046) 0.263 (p=0.063)
7:    RF (biomarkers)    4 0.444 (p=0.704) 0.275 (p=0.759)
8:                       8 0.462 (p=0.595) 0.274 (p=0.577)
9:                      12 0.573 (p=0.119) 0.214 (p=0.086)
```

Sensitivity analysis: Similar results were obtained with the complete case analysis (which also drop 3 biomarkers):

```
                model week             AUC           Brier
1: GLM (no biomarker)    4 0.501 (p=0.210) 0.248 (p=0.308)
2:                       8  0.56 (p=0.062) 0.255 (p=0.159)
3:                      12 0.601 (p=0.032) 0.206 (p=0.038)
4:   GLM (biomarkers)    4 0.443 (p=0.656) 0.284 (p=0.626)
5:                       8 0.637 (p=0.025) 0.253 (p=0.022)
6:                      12 0.585 (p=0.133)  0.23 (p=0.119)
7:    RF (biomarkers)    4 0.491 (p=0.444) 0.261 (p=0.466)
8:                       8 0.537 (p=0.246) 0.261 (p=0.221)
9:                      12 0.689 (p=0.011) 0.194 (p=0.016)
```

The main difference being at week 12 where the RF results appeared better (both in term of AUC and brier score) and the logistic model worse.

IMPORTANT NOTE: what is missing here is a test comparing `GLM (no biomarker)` to the other 2 models. I'm not sure yet how to do that correctly.
Also some p-values are a bit weird e.g. AUC=0.501 with p=0.210. I'll double check.

4 Conclusion

There is some evidence that cognition and EEG (and to a lesser extend OFC thickness) are predictive of recovery at week 8. By some evidence, we mean that the unadjusted p-value was significant (between 0.01 and 0.05) while the adjusted p-value was above the traditional threshold (typically around 0.1). There was also some evidence for a predictive value of these biomarkers: performance superior to the null predictor and pointwise estimate of the in gain in AUC when adding the biomarkers of about +0.13. However the Brier score was quite high and not different from the null predictor. So while the biomarkers may help to discriminate between patients who will recover or not, it seems that we are not able to obtain reliable (nor clinically relevant) probabilities of recovery.
Using a flexible model such as random forest did not seems to help, which is to be expected with a rather limited sample size.

Time at which recovery was assessed also appeared to have a impact on the results. For instance the value of EEG was only evident at week 8. Cognition seemed to have a more stable association with recovery while others like age or OFC thickness the strength of assocation seemed to increase over time. Overall, there was no evidence that the biomarkers were useful at week 4 while there was some evidence at week 4, probably due to cognition and EEG, and in the complete case analysis also some evidence at week 12, probably due to OFC thickness and PET.

5 References

Goerigk, S. A., Padberg, F., Bühner, M., Sarubin, N., Kaster, T. S., Daskalakis, Z. J., Blumberger, D. M., Borrione, L., Razza, L. B., and Brunoni, A. R. (2021). Distinct trajectories of response to prefrontal tdcs in major depression: results from a 3-arm randomized controlled trial. *Neuropsychopharmacology*, 46(4):774–782.

Moons, K. G., Donders, R. A., Stijnen, T., and Harrell Jr, F. E. (2006). Using the outcome for imputation of missing predictor values was preferred. *Journal of clinical epidemiology*, 59(10):1092–1101.

Van Buuren, S. and Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in r. *Journal of statistical software*, 45:1–67.

Appendix A ROC curves

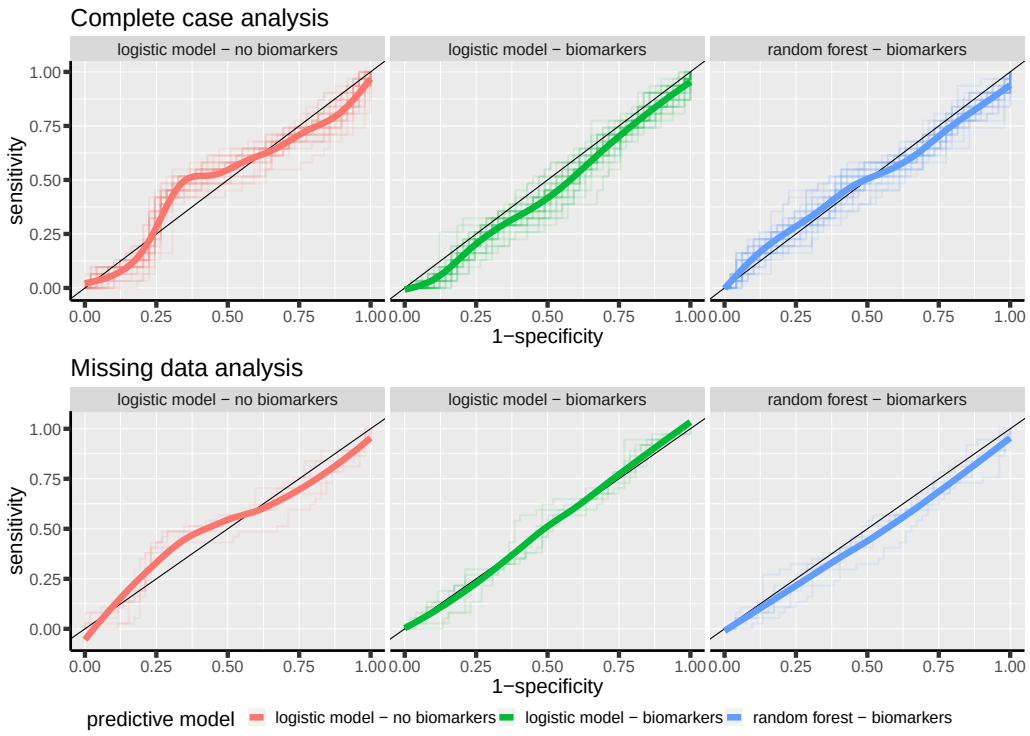

Figure A: Week 4

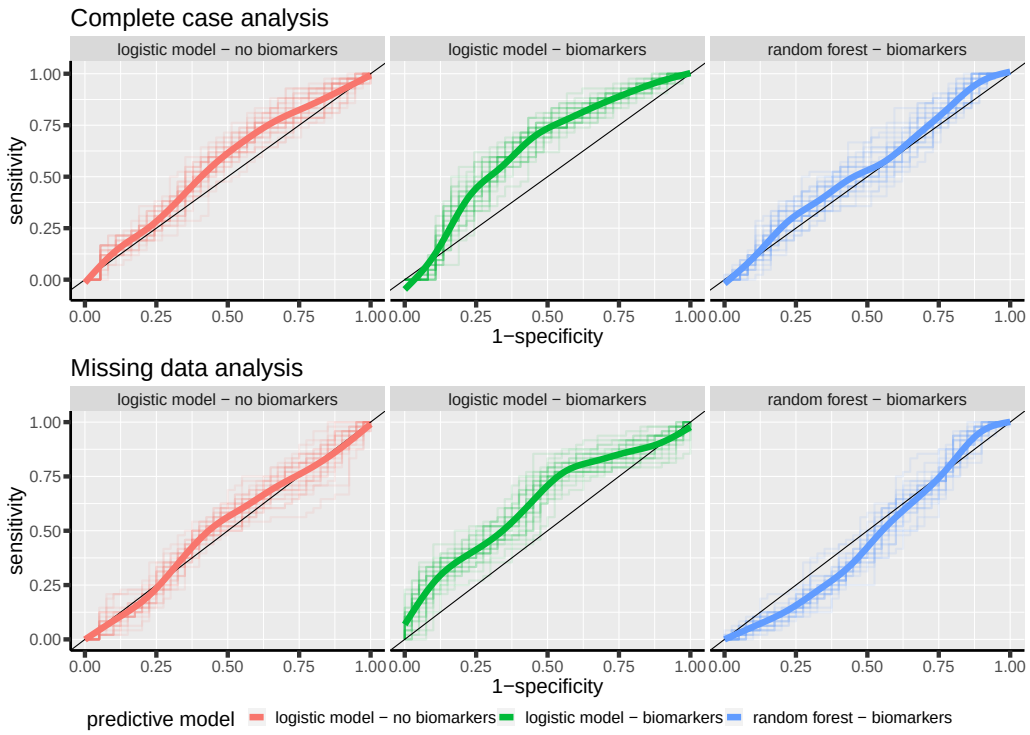

Figure B: Week 8

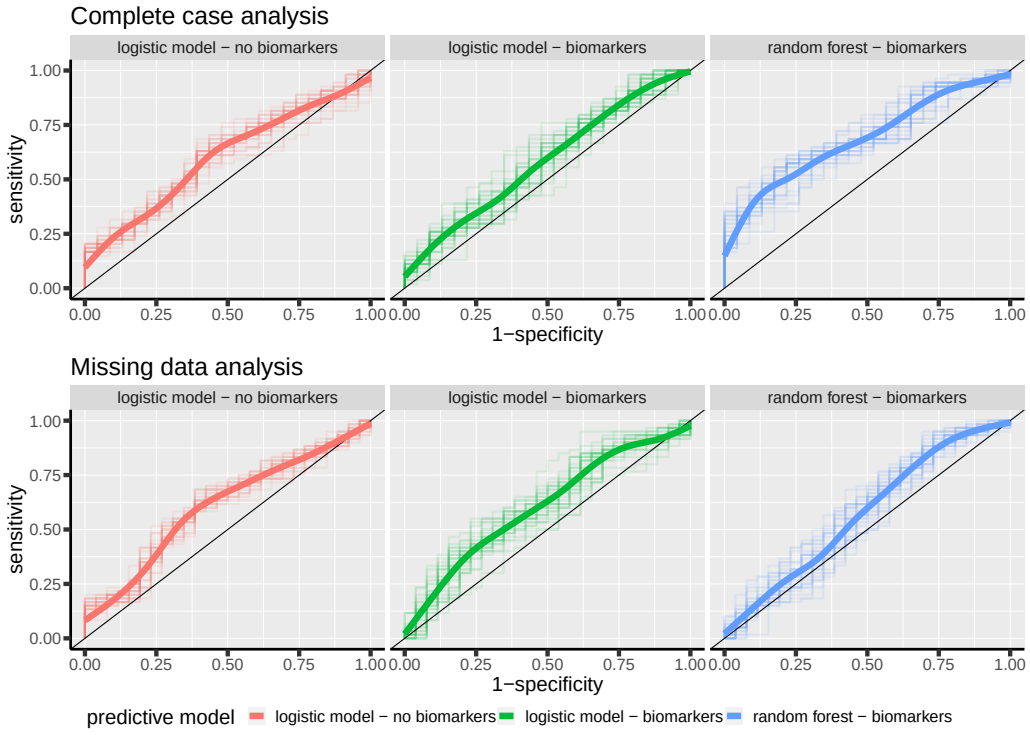

Figure C: Week 12