Supplementary material for "Preoperative Cognitive Profile Predictive of Cognitive Decline after Subthalamic Deep Brain Stimulation in Parkinson's Disease"

Josef Mana[1], Ondrej Bezdicek[1], Andrej Lasica[1], Filip Ruzicka[1], Anna Fecikova[1], Olga Klempirova[1], Tomas Nikolai[1], Tereza Uhrova[1], Evzen Ruzicka[1], Dusan Urgosik[2], and Robert Jech [1]

[1]Department of Neurology and Centre of Clinical Neuroscience, First Faculty of Medicine and General University Hospital in Prague, Charles University, Czech Republic

[2]Department of stereotactic and radiation neurosurgery, Na Homolce Hospital, Prague, Czech Republic

Author Note

Correspondence concerning this article should be addressed to Josef Mana, Email: josef.mana@protonmail.com

Supplementary material for "Preoperative Cognitive Profile Predictive of Cognitive Decline after Subthalamic Deep Brain Stimulation in Parkinson's Disease"

In this supplementary material we present additional information to the article *"Preoperative Cognitive Profile Predictive of Cognitive Decline after Subthalamic Deep Brain Stimulation in Parkinson's Disease"* including further presentation of results that was not included in the main text due to space constraints. All procedures described in this supplementary material are accompanied by R code used to implement the steps described herein and Stan code for Bayesian generalized linear mixed models (GLMMs) fitted during this project. The R code and Stan models as well as raw files containing all images and tables are available at `https://github.com/josefmana/dbs_cogPRED`. Since the data used for model fitting in our study contain medical records of included patients, they are not publicly available for privacy reasons. Moreover, because GLMMs reported in this article are exceedingly large for purposes of online storage (> 2 GB each), only the R and Stan codes are included.

Pre-surgery cross-sectional exploratory factor analysis

Data pre-processing

For exploratory factor analyses (EFAs) we first log transformed all response time-based tasks (i.e., Trail Making Test and Stroop test), then standardized (i.e., mean-centered and scaled by their in-sample standard deviation) all variables before applying multiple imputations for missing values. EFA was then fitted on each imputed data set via ordinary least squares to find the minimal residual solution. This procedure was repeated for three up to eight factor solutions.

Supplementary presentation of results

Supplementary EFA results are presented in Table 1 and Figure 1 (see below). Table 1 presents numerical summary of fit indexes of each three to eight factor solutions across one hundred imputations. Note that Tucker-Lewis Index (TLI) was above the threshold implying good fit (TLI > 0.9) in only three **fourths of** six-factor models, but it was above this threshold in all but three out of one hundred seven-factor models.

Similar information is visually presented in Figure 1 which depicts density plots of TLI and upper 90% confidence interval bound of root-mean-square-error approximation (RMSEA) of all models across imputations. This clear improvement in fit of seven- compared to six-factor model, only modest improvement of eight- compared to seven-factor model, and overall theoretical plausibility of factors identified by the seven-factor model led us to retain seven factors for further analyses.

Table 1

Summary of fit indexes of the exploratory factor analysis across one datasets

Model	TLI	RMSEA	RMSEA 90% CI (upper bound)	Total variance accounted for
3-factor	0.68 (0.03)	0.09 (0.00)	0.11 (0.00)	0.38 (0.01)
4-factor	0.81 (0.03)	0.07 (0.01)	0.09 (0.00)	0.44 (0.01)
5-factor	0.87 (0.03)	0.06 (0.01)	0.08 (0.01)	0.48 (0.01)
6-factor	0.92 (0.03)	0.04 (0.01)	0.07 (0.01)	0.52 (0.01)
7-factor	0.96 (0.03)	0.03 (0.01)	0.06 (0.01)	0.55 (0.01)
8-factor	0.99 (0.03)	0.02 (0.01)	0.05 (0.01)	0.58 (0.01)

Values represent mean (SD) or percentages if indicated in brackets.

TLI Tucker-Lewis Index. RMSEA root-mean-square-error approximation. CI confidence interval

Longitudinal generalized linear mixed models

Data pre-processing

To simplify the process of choosing appropriate prior distributions and minimize multicollinearity, all variables were standardized (i.e., mean-centered and scaled by their in-sample standard deviation) before the analyses. The only variable that was not pre-processed this way was time after surgery. This variable was entered into all models on its raw scale (i.e., years after surgery) shifted forward by a median time of pre-surgery assessment (i.e., 0.30 years). Consequently, model intercepts represent estimates of patients' cognitive performance in Mattis Dementia Rating Scale (DRS-2)

at pre-surgery assessment (0.30 years before surgery) and time slopes represent DRS-2 annual post-surgery cognitive decline. Before they were entered into the models, all pre-surgery cognitive factors and test scores were coded such that higher values indicated poorer performance. Parameters associated with these variables (see Figure 5, Figure 6, Table 2, Table 3 as well as Figure 4 in the main text) thus represent an effect of a (relative) pre-surgery deficit in **the** corresponding latent cognitive factor or manifest cognitive test score on prediction of pre-surgery DRS-2 (the β parameters) and post-surgery annual decline in DRS-2 (the δ parameters). Negative parameter values imply that pre-surgery cognitive deficit unfavorably affects the outcome and vice versa for positive parameter values.

Posterior predictive check

To validate in-sample fit of our predictive models, we computed models' "predictions" for each included patient and compared these predictions to observed values (see Figure 2). Note that since one of the advantages of multilevel modelling is partial pooling, i.e., shrinking parameter estimates towards each other and thus down-weighting the effect of influential outliers to reduce overfitting, models **are** neither expected nor required to replicate observed values exactly. Our models show reasonable fit to most patients with clear shrinkage in case of outliers (for instance patient S045 in Figure 2). Furthermore, while the **"time only,"** "test scores" and the "factor scores" model provide similar posterior predictions for our patients, the "**factor** scores" model was influenced by outlying values to a small**er** degree **compared to the "time only" and "test scores" models** (for instance patients S023, S107 or S124).

Supplementary presentation of results

In Figure 3 we present visual summary of statistical estimates for both sample and population versions of our *RQ1* estimand (see Table 4 in the main text for numerical summary of this figure) as well as models prediction of raw data yet unobserved. For this presentation we selected 90% posterior equal-tailed intervals (ETIs) so that there is 5% probability the true score difference falls below the bottom bound and 5% probability

the true score difference falls above the top bound of the interval. Notice that estimates for the sample are significantly more certain than estimates generalized to the assumed population via adding patient-level variability.

In Figure 4 we present side-by-side comparison of patient-level true score variability that remains after accounting for association between time after surgery and cognitive performance (of the descriptive time only model) compared to patient-level true score variability after accounting for association between cognitive performance and time after surgery combined with pre-surgery cognitive profile (of the predictive test scores and factor scores models). To ensure estimates are comparable across models, we refitted the time only model using regularizing (Lasso) priors identical to predictive models' priors instead of the non- and weakly-informative default *brms* priors of the time only model presented in the main text. Although it can be seen that regularizing priors shifted posterior distributions of patient-specific parameters towards zero (compare the time only (default) and time only (Lasso) estimates in Figure 4), adding pre-surgery cognitive profile as predictor shifted these distributions even further (compare the test and factor scores to the time only (Lasso) estimates in Figure 4).

In Table 2 we present numerical summary of group-level posterior parameters of the "test scores" model while in Table 3 we present numerical summary of group-level posterior parameters of the "factor scores" model which supplement the information presented in Figure 3 in the main text. Since only the interaction terms (i.e., the δ parameters) comprised empirical estimands for our query (*RQ2*), the remaining parameters were omitted from the main text.

Table 2

Summary of group-level effects' posteriors from the "test scores" generalized linear mixed model reported in the main text

Parameter	95% PPI

Global intercept ()			
Intercept	140.17	[139.58, 140.75]	0.000
Baseline correlates ()			
TMT-A	0.00	[-0.35, 0.38]	0.486
TMT-B	-0.26	[-0.87, 0.16]	0.884
DS-F	-0.05	[-0.50, 0.28]	0.655
DS-B	-0.07	[-0.54, 0.26]	0.699
LNS	-0.19	[-0.77, 0.18]	0.844
SS-F	-0.05	[-0.51, 0.29]	0.654
SS-B	-0.10	[-0.57, 0.23]	0.741
TOL	-0.07	[-0.53, 0.26]	0.700
PST-D	0.03	[-0.31, 0.47]	0.394
PST-W	0.00	[-0.39, 0.40]	0.499
PST-C	-0.35	[-0.94, 0.10]	0.935
COWAT	0.00	[-0.36, 0.37]	0.498
CFT	-0.15	[-0.72, 0.21]	0.801
Sim.	-0.19	[-0.76, 0.16]	0.848
RAVLT-IR	-0.06	[-0.51, 0.29]	0.660
RAVLT-B	-0.35	[-0.96, 0.10]	0.936
RAVLT-DR	0.06	[-0.28, 0.54]	0.330
RAVLT-Rec50	-0.01	[-0.39, 0.35]	0.537
RAVLT-Rec15	-0.11	[-0.59, 0.21]	0.769
FP-IR	-0.06	[-0.54, 0.30]	0.669
FP-DR	-0.04	[-0.52, 0.34]	0.617
STAI-X1	-0.00	[-0.36, 0.36]	0.509
STAI-X2	0.01	[-0.34, 0.38]	0.472
Time-dependent effects ()			
Time	-0.72	[-0.98, -0.47]	1.000

TMT-A Œ Time	-0.08	[-0.32, 0.11]	0.810
TMT-B Œ Time	-0.15	[-0.45, 0.09]	0.887
DS-F Œ Time	0.12	[-0.08, 0.35]	0.130
DS-B Œ Time	0.07	[-0.14, 0.32]	0.242
LNS Œ Time	0.07	[-0.15, 0.33]	0.255
SS-F Œ Time	0.23	[-0.06, 0.57]	0.057
SS-B Œ Time	-0.11	[-0.39, 0.11]	0.838
TOL Œ Time	-0.05	[-0.27, 0.15]	0.696
PST-D Œ Time	-0.01	[-0.27, 0.23]	0.561
PST-W Œ Time	-0.15	[-0.44, 0.08]	0.896
PST-C Œ Time	-0.09	[-0.35, 0.12]	0.811
COWAT Œ Time	-0.14	[-0.36, 0.05]	0.922
CFT Œ Time	-0.02	[-0.24, 0.20]	0.581
Sim. Œ Time	0.08	[-0.13, 0.34]	0.227
RAVLT-IR Œ Time	0.00	[-0.23, 0.24]	0.478
RAVLT-B Œ Time	0.02	[-0.16, 0.24]	0.392
RAVLT-DR Œ Time	0.07	[-0.13, 0.31]	0.228
RAVLT-Rec50 Œ Time	-0.03	[-0.27, 0.17]	0.632
RAVLT-Rec15 Œ Time	-0.00	[-0.22, 0.21]	0.521
FP-IR Œ Time	-0.03	[-0.34, 0.26]	0.594
FP-DR Œ Time	-0.05	[-0.39, 0.22]	0.684
STAI-X1 Œ Time	-0.00	[-0.20, 0.18]	0.521
STAI-X2 Œ Time	-0.00	[-0.21, 0.20]	0.510

All cognitive predictors were scaled such that negative values mean negative effect of pre-surgery deficit on longitudinal cognitive trajectory.

b: parameter value point estimate (posterior median); PPI: posterior probability interval; Pr(b < 0): probability that a parameter is negative, i.e., probability that the predictor has a negative effect on the outcome (this quantity does not apply to Intercept where it cannot be interpreted but it is reported for completeness); Œ: statistical interaction term; STAI-X1: State-Trait Anxiety Inventory, the state version; STAI-X2: State-Trait Anxiety Inventory, the trait version; TMT-A: Trail Making Test, part A; TMT-B: Trail Making Test, part B; DS-F: Digit Span forward; DS-B: Digit Span backward; LNS: letter-number sequencing; SS-F: Spatial Span forward; SS-B: Spatial Span backward; TOL: Tower of London task; PST-D: Prague Stroop Test, dot color naming; PST-W: Prague Stroop Test, word color naming; PST-C: Prague Stroop Test, interference condition; COWAT: Controlled Oral Word Association Test; CFT: category fluency test; Sim.: Similarities; RAVLT-IR: Rey Auditory Verbal Learning Test, immediate recall; RAVLT-B: Rey Auditory Verbal Learning Test, recall of the interference set; RAVLT-DR: Rey Auditory Verbal Learning Test, delayed recall; RAVLT-Rec50: Rey Auditory Verbal Learning Test, delayed recognition from 50 items (15 correct answers + 35 distractors); RAVLT-Rec15: Rey Auditory Verbal Learning Test, delayed recognition, number of correctly identified from 15 items; FP-IR: Family Pictures, immediate recall; FP-DR: Family Pictures, delayed recall.

Table 3

Summary of group-level effects' posteriors from the "factor scores" generalized linear mixed model reported in the main text

Parameter		95% PPI	
Global intercept ()			
Intercept	140.25	[139.67, 140.83]	0.000
Baseline correlates ()			

EF/Att.	-0.19	[-0.78, 0.28]	0.790
EM	-0.17	[-0.71, 0.26]	0.787
VWM	-0.92	[-1.68, -0.11]	0.991
VM	-0.35	[-1.02, 0.19]	0.889
SS	-0.73	[-1.39, -0.03]	0.985
An.	-0.06	[-0.59, 0.40]	0.613
SWM	-0.32	[-1.05, 0.23]	0.861
Time-dependent effects ()			
Time	-0.75	[-0.99, -0.51]	1.000
EF/Att. Œ Time	-0.39	[-0.63, -0.15]	0.999
EM Œ Time	-0.00	[-0.22, 0.22]	0.510
VWM Œ Time	0.17	[-0.09, 0.45]	0.099
VM Œ Time	-0.17	[-0.44, 0.10]	0.888
SS Œ Time	-0.14	[-0.47, 0.18]	0.779
An. Œ Time	-0.00	[-0.21, 0.21]	0.504
SWM Œ Time	0.06	[-0.33, 0.41]	0.367

All cognitive predictors were scaled such that negative values mean negative effect of pre-surgery deficit on longitudinal cognitive trajectory.

b: parameter value point estimate (posterior median); PPI: posterior probability interval; Pr(b $<$ 0): probability that a parameter is negative, i.e., probability that the predictor has a negative effect on the outcome (this quantity does not apply to Intercept where it cannot be interpreted but it is reported for completeness); Œ: statistical interaction term; EF/Att.: Executive functions/Attention; EM: Episodic memory; VWM: Verbal working memory; VM: Visuospatial memory; SS: Set shifting; An: Anxiety; SWM: Spatial working memory.

Robustness checks

To confirm that our results are robust to effects of aging, dopaminergic medication, and depressive symptoms, we carried out a robustness check by fitting

parallel "test scores" and "factor scores" models with additional group-level predictors age, levodopa equivalent daily dose (LEDD) and Beck Depression Inventory (BDI-II) **as well as their interaction with time after surgery**. Stan code for each of these models with the exact specification is available at https://github.com/josefmana/dbs_cogPRED and is equivalent to the models reported in the main text.

Side-to-side comparison of models' group-level parameters' posterior summaries is presented in Figure 5 for the "test scores" model and in Figure 6 for the "factor scores" model. All models arrived at similar posteriors implying our results are robust to effects of aging, dopaminergic medication, and depressive symptoms. Importantly, the empirical estimands relating to *RQ2* (i.e., the time-dependent effects represented by the δ parameters) are similar across models leading to identical substantive conclusions.

Exploring electrode localization

To explore association between electrode localization and post-surgery cognitive decline, we retrospectively gathered magnetic resonance imaging (MRI) data of patients from our data set and estimated volume of affected tissue (VAT) based on stimulation parameters at the time of MRI assessment and its intersection with subthalamic nucleus (STN) as well as its motor, associative, and limbic components.

Only patients with all (i) pre-surgery MRI for STN localization, (ii) post-surgery MRI for electrode localization, and (iii) stimulation parameters at time of MRI assessment for VAT computation were included. Pre-surgery images included T1-weighted (T1w) MPRAGE sequence acquired on 3 Tesla Siemens Symphony, Skyra or TrioTim System (Siemens, Erlangen, Germany) with Echo Time (TE) = 2.43-4.43 ms, Repetition Time (TR) = 1,870-2,300 ms, Flip Angle (FA) = 8-15°, and Slice Thickness = 1-1.5 mm varying slightly across patients. Furthermore, pre-surgery T2-weighted (T2w) Spin echo scans were acquired on the same MRI machine as pre-surgery T1w with TE = 80-90 ms, TR = 2,000-2,440 ms, FA = 90°, and

Slice Thickness = 2 mm. Post-surgery, T1w MPRAGE sequence was acquired on 1.5 Tesla Siemens Symphony, Skyra or Avanto System with TE = 2.43-3.93 ms, TR = 2,060-2,200 ms, FA = 8-15°, and Slice Thickness = 0.9-1.0 mm. Pre-processing was conducted in Lead-DBS software (lead-dbs.org). The outcome of this pre-processing was a set of estimated overlaps between VAT and STN (and its components) for each included patient.

Data set

From total of 126 patients 69 were included into analysis, 38 were excluded due to missing MRI at both pre- and post-surgery assessments, 18 patients were excluded due to missing post-surgery MRI and 1 patient was excluded due to missing stimulation parameters at time of MRI. While according to our estimates, there was higher mean overlap of VATs and motor STN compared to other STN components, for some patient we estimated high proportion overlap between their VAT and associative STN as well (see Table 4). Moreover, our estimates imply that there was only small or no overlap between VAT and at least one STN in appreciable number of patients (see also Figure 7). Since all patients in our data set showed at least some clinical improvement after STN DBS according to attending physicians, these cases can be attributed to noisy estimates rather than true misalignment between the lead and STN.

Table 4

Overlap between volume of activated tissue and subthalamic nucleus components

Hemisphere	N	Md	Min-Max	M	SD
Subthalamic nucleus (STN)					
Left	67	4.15	0.00-50.31	7.75	9.61
Right	68	3.24	0.00-46.40	6.27	8.78

Motor STN					
Left	67	6.56	0.00-49.64	9.29	10.89
Right	68	5.30	0.00-65.83	8.88	12.49
Associative STN					
Left	67	1.66	0.00-71.61	7.29	12.38
Right	68	1.14	0.00-50.46	5.42	9.46
Limbic STN					
Left	67	2.36	0.00-49.11	4.44	7.08
Right	68	1.31	0.00-28.89	4.09	6.69

N: number of observations; Md: median; M: mean; SD: standard deviation; Numbers in all columns but N represent percentage points of overlap between estimated volume of activated tissue and patient's subthalamic nucleus.

Statistical models

Since we were able to use only subset of patients for analysis of association between electrode localisation and post-surgery cognitive decline, we first directly compared included versus excluded patients' clinical and neuropsychological baseline characteristic. We applied "Bayesian t-tests" for difference in means with unequal standard deviations of the following form:

$$DV_i \sim \ N(\mu_0, \sigma_0), for\ excluded\ patients$$

$$DV_j \sim \ N(\mu_1, \sigma_1), for\ included\ patients$$

$$\mu_x \sim \ N(0,1), forx \in \{0,1\}$$

$$\sigma_x \sim \ E(1), forx \in \{0,1\}$$

where $i = 1 \ldots n_0$ with n_0 excluded patients, $j = 1 \ldots n_1$ with n_1 included patients, *DV* is a dependent variable (i.e., the clinical or neuropsychological characteristic in question) *N()* is the Normal probability density function and *E()* is the Exponential probability density function. All

dependent variables were standardized (i.e., mean-centered and scaled by their in-sample standard deviation) and response times log-transformed and then standardized before entering the analysis. The between-group differences were evaluated by calculating difference scores $\mu_1 - \mu_0$ for means and $\sigma_1 - \sigma_0$ for standard deviations. Posterior medians ad 95% PPI of these difference scores were then reported.

In the next step we directly compared post-surgery cognitive decline of included versus excluded patients. To do this, we adjusted the descriptive model from the main text to include slopes varying by inclusion/exclusion status:

$$P(DRS_i = DRS_{max}) = 1 - T(\vartheta, \mu_i, \sigma), for DRS_i \in N_{max}, N_{max} = \{i : drs_i = drs_{max}\}$$

$$DRS_i \sim \ t(\vartheta, \mu_i, \sigma), for DRS_i \in N_1, N_1 = \{i : drs_i < drs_{max}\}$$

$$\mu_i = \alpha + \delta_{time} time_i + \beta_{inclusion} inclusion_i + \delta_{inclusion} time_i inclusion_i + \tau_{\bar{\alpha}} z_{\bar{\alpha},id[i]} + \tau_{\bar{\delta}} z_{\bar{\delta},id[i]} time_i$$

with default brms priors for all parameters (see Stan file at https://github.com/josefmana/dbs_cogPRED for more information). We then extracted a measure of cognitive decline discrepancy between the two subsamples by re-scaling $\delta_{inclusion}$ to the original DRS-2 points/year scale.

Finally, to evaluate predictive value of VATs overlap with STN components, we adjusted the predictive models from the main text:

$$P(DRS_i = DRS_{max}) = 1 - T(\vartheta, \mu_i, \sigma), for DRS_i \in N_{max}, N_{max} = \{i : drs_i = drs_{max}\}$$

$$DRS_i \sim \ t(\vartheta, \mu_i, \sigma), for DRS_i \in N_1, N_1 = \{i : drs_i < drs_{max}\}$$

$$\mu_i = \alpha + \delta_{time} time_i + \sum_{j=1}^{m} (\beta_{predictor[j]} predictor_{[j]i} + \delta_{predictor[j]} time_i predictor_{[j]i}) + \tau_{\bar{\alpha}} z_{\bar{\alpha},id[i]} + \tau_{\bar{\delta}} z_{\bar{\delta},id[i]} time_i$$

such that each $predictor_{[j]i}$ represents one of the six combinations of VAT intersection with STN component (motor, associative and limbic) separately for each hemisphere. Furthermore, to reduce the amount of

regularization we used *N(0,0.5)* priors for population level parameters α, δ_{time}, $\beta_{predictor[j]}$ and $\delta_{predictor[j]}$ instead of Bayesian Lasso. All variables were standardized (i.e., mean-centered and scaled by their in-sample standard deviation) befored entering the analysis. We used the set of $\delta_{predictor[j]}$ values as a measure representing the expected prognostic value of DBS affected proportion of STN for post-surgery cognitive decline rate. All models described in this section were fitted using four chains each with 2,500 samples out of which 500 samples were discarded as a warm-up.

Results

Table 5

Mean differences between included and excluded patients in pre-surgery c
neuropsychological variables

	Descriptive statistics			
	N	Excluded[1]	Included[1]	
Age (years)	57/69	58.42 ś 7.28	55.65 ś 8.23	-2.74 [-5.45, -0.
Disease duration (years)	56/69	11.86 ś 4.47	10.99 ś 3.83	-0.87 [-2.37, 0.6
Education (years)	48/69	13.77 ś 2.50	14.59 ś 3.14	0.81 [-0.21, 1.8
Sex (males)	57 69	38 (30.2%)	45 (35.7%)	-
Age at surgery (years)	57/69	58.70 ś 7.35	56.04 ś 8.28	-2.61 [-5.35, 0.1
Disease duration at surgery (years)	56/69	12.02 ś 4.40	11.39 ś 3.75	-0.60 [-2.01, 0.9
LEDD (mg)	49/65	1646.17 ś 682.46	1735.12 ś 667.34	87.50 [-159.72, 33
Levodopa test (% response)	32/61	55.20 ś 13.41	51.29 ś 12.39	-3.79 [-9.52, 1.8
MDS-UPDRS III (ON medication)	41/64	20.27 ś 6.43	22.75 ś 8.12	2.43 [-0.28, 5.2
MDS-UPDRS III (OFF medication)	38/62	44.54 ś 9.63	46.56 ś 11.66	1.96 [-2.10, 6.4
DRS-2 (range 0-144)	57/69	140.12 ś 4.03	139.48 ś 3.37	-0.63 [-1.99, 0.6
BDI-II (range 0-63)	55/67	10.84 ś 6.03	8.00 ś 5.61	-2.79 [-4.91, -0.
STAI-X1 (range 20-80)	41/63	41.49 ś 8.98	36.17 ś 7.83	-5.21 [-8.48, -1.

STAI-X2 (range 20-80)	41/63	41.63 ś 6.89	38.14 ś 8.59	-3.45 [-6.43, -0.
TMT-A (secs)	57/68	47.39 ś 19.23	39.60 ś 11.31	-5.99 [-11.07, -1
TMT-B (secs)	56/68	122.41 ś 56.43	116.21 ś 53.99	-3.05 [-19.58, 12
DS-F (range 0-16)	44/69	8.91 ś 1.94	8.96 ś 2.08	0.04 [-0.74, 0.8
DS-B (range 0-14)	44/69	6.30 ś 1.86	6.16 ś 1.77	-0.13 [-0.86, 0.5
LNS (range 0-21)	34/63	7.74 ś 2.27	7.90 ś 2.57	0.16 [-0.84, 1.1
SS-F (range 0-16)	43/67	7.58 ś 1.56	7.51 ś 1.86	-0.07 [-0.74, 0.5
SS-B (range 0-16)	43/67	6.74 ś 1.53	7.12 ś 1.78	0.36 [-0.29, 0.9
TOL (range 0-108)	53/65	71.89 ś 10.74	77.42 ś 8.27	5.43 [1.96, 9.1
PST-D (secs)	56/68	13.16 ś 2.10	13.04 ś 2.59	-0.18 [-1.02, 0.5
PST-W (secs)	56/68	16.14 ś 3.28	15.38 ś 2.66	-0.65 [-1.74, 0.3
PST-C (secs)	56/68	31.27 ś 10.01	27.78 ś 8.11	-2.97 [-6.20, 0.0
COWAT (total words)	57/68	31.18 ś 8.95	33.34 ś 9.08	2.12 [-1.06, 5.3
CFT (words/min.)	29/60	19.24 ś 6.88	24.15 ś 6.70	4.80 [1.55, 7.6
Sim. (range 0-28)	32/62	22.50 ś 3.64	21.15 ś 4.63	-1.31 [-3.05, 0.4
RAVLT-IR (range 0-75)	39/69	43.10 ś 8.96	44.19 ś 8.10	1.07 [-2.37, 4.4
RAVLT-B (range 0-15)	39/69	4.87 ś 1.54	4.62 ś 1.40	-0.24 [-0.84, 0.3
RAVLT-DR (range 0-15)	39/69	8.28 ś 2.20	8.42 ś 2.65	0.14 [-0.82, 1.0
RAVLT-Rec50 (range 0-50)	38/67	44.82 ś 4.01	45.25 ś 3.19	0.43 [-1.00, 1.9
RAVLT-Rec15 (range (0-15)	38/69	13.24 ś 1.62	13.36 ś 1.50	0.12 [-0.51, 0.7
FP-IR (range 0-64)	48/26	32.56 ś 11.08	31.08 ś 8.49	-1.47 [-5.99, 3.0
FP-DR (range 0-64)	48/26	32.33 ś 10.86	31.12 ś 8.23	-1.17 [-5.42, 3.4

[1]Values indicate mean ś standard deviation or frequency (percentage)

N: number of observations from excluded/included patients; : inference for mean difference between groups; : estimate of difference between standard deviations between groups; d: difference value point estimate (posterior median) [95% posterior probability interval] Pr(d $<$ 0): probability that a difference is negative, i.e., probability that included patients had higher mean/stnadard deviation than excluded patients; MDS-UPDRS III: Movement Disorder Society Unified Parkinsons Disease Rating Scale, motor part; LEDD: levodopa equivalent daily dose; Levodopa test: a percentage change of the MDS-UPDRS III score from medication OFF to medication ON state during the levodopa test; STAI-X1: State-Trait Anxiety Inventory, the state version; STAI-X2: State-Trait Anxiety Inventory, the trait version; TMT-A: Trail Making Test, part A; TMT-B: Trail Making Test, part B; DS-F: Digit Span forward; DS-B: Digit Span backward; LNS: letter-number sequencing; SS-F: Spatial Span forward; SS-B: Spatial Span backward; TOL: Tower of London task; PST-D: Prague Stroop Test, dot color naming; PST-W: Prague Stroop Test, word color naming; PST-C: Prague Stroop Test, interference condition; COWAT: Controlled Oral Word Association Test; CFT: category fluency test; Sim.: Similarities; RAVLT-IR: Rey Auditory Verbal Learning Test, immediate recall; RAVLT-B: Rey Auditory Verbal Learning Test, recall of the interference set; RAVLT-DR: Rey Auditory Verbal Learning Test, delayed recall; RAVLT-Rec50: Rey Auditory Verbal Learning Test, delayed recognition from 50 items (15 correct answers + 35 distractors); RAVLT-Rec15: Rey Auditory Verbal Learning Test, delayed recognition, number of correctly identified from 15 items; FP-IR: Family Pictures, immediate recall; FP-DR: Family Pictures, delayed recall.

Pre-surgery differences between included and excluded patients are presented in Table 5 and Figure 8. The descriptive model comparing post-surgery cognitive decline in excluded versus included patients had satisfactory convergence statistics ($\hat{R}s \leq 1.017$). All observations had Pareto-k below 0.60. According to the model, included patients experienced an average post-surgery decline of 0.70 DRS-2 points/year (95% PPI [-1.10,

-0.30]) from an average pre-surgery DRS-2 performance of 139.79 points (95% PPI [138.83, 140.70]) whereas excluded patients experienced an average post-surgery decline of 1.16 DRS-2 points/year (95% PPI [-1.62, -0.73]) from an average pre-surgery DRS-2 performance of 141.08 points (95% PPI [140.02, 142.25]). Although, on average the post-surgery cognitive decline was slower in included patients by 0.46, the difference was not statistically clear based on 95% PPI [-0.15, 1.05].

Finally, the predictive model comparing post-surgery cognitive decline depending on proportion of STN components volume being affected by VAT had satisfactory convergence statistics ($\hat{R}s \leq 1.003$). However, there were 11 potentially influential observations with Pareto-k above 0.7, the highest Pareto-k reached the value of 0.90. Results summarising group-level effects are presented in Figure 9, Figure 10, and Table 6.

Table 6

Summary of group-level effects' posteriors from the generalized linear mixed model predicting cognitive performance by subthalamic nucleus components being stimulated

Parameter		95% PPI	
Global intercept ()			
Intercept	139.75	[138.89, 140.62]	0.000
Baseline correlates ()			
ML	-0.07	[-1.27, 1.01]	0.552
MR	-0.31	[-1.38, 0.79]	0.706
AL	-0.62	[-2.08, 0.80]	0.802
AR	-0.86	[-2.55, 0.82]	0.850
LL	1.47	[-1.53, 4.54]	0.171
LR	0.60	[-2.14, 3.31]	0.327
Time-dependent effects ()			

Time	-0.77	[-1.18, -0.36]	1.000
ML Œ Time	0.01	[-0.53, 0.55]	0.479
MR Œ Time	-0.01	[-0.45, 0.43]	0.511
AL Œ Time	0.12	[-0.50, 0.72]	0.347
AR Œ Time	-0.29	[-1.01, 0.40]	0.796
LL Œ Time	0.11	[-1.64, 1.70]	0.448
LR Œ Time	0.14	[-1.01, 1.37]	0.415

Parameters represent expected increase/decrease of cognitive performance assessed via Dementia Rating Scale, second edition, associated with observing an increase of volume of activated tissue (VAT) intersection proportion with subthalamic nucleus (STN) component by 10 % of total volume of the component. Time dependent effects measure association between predictor and outcome per one year.

b: parameter value point estimate (posterior median); PPI: posterior probability interval; Pr(b $<$ 0): probability that a parameter is negative, i.e., probability that the predictor has a negative effect on the outcome (this quantity does not apply to Intercept where it cannot be interpreted but it is reported for completeness); Œ: statistical interaction term; ML: left motor STN; MR: right motor STN; AL: left associative STN; AR: right associative STN; LL: left limbic STN; LR: right limbic STN.

Estimation of false positive rates

To test our assumption that the classical two-step **univariable screening** procedure of identifying significant pre-surgery predictors of post-surgery cognitive decline leads to inflated false positive rates that can be alleviated by either dimension reduction of the matrix of predictors (e.g., via factor analysis) or applying the Bayesian Lasso, we conducted a series of simulations **based** on the data structure equivalent to our data set with respect to the number of patients, the number of observations per patient, time from surgery of each observation and the number of potential predictors. Simulated data sets used to produce our results as well as generating functions are available at `https://github.com/josefmana/dbs_cogPRED`. The reader should navigate to the **"scripts/sims.R"** file to validate our findings and check their

robustness to change in parameters which were omitted from current study for parsimony and computational time reasons.

Data-generating process

The outcome (representing idealized cognitive screening score) followed Gaussian distribution with unit variation and mean shifted from zero by (i) patient-specific intercept (pre-surgery shift) and slope (post-surgery decline shift) and (ii) an average annual post-surgery decline (i.e., population-level slope denoted b_1 henceforth). The b_1 parameter was set to either no ($b_1 = 0$), mild ($b_1 = -0.3$) or moderate ($b_1 = -0.5$) average annual post-surgery decline (see columns of Figure 12). Moreover, total of seven or twenty-three potential pre-surgery predictors were generated for each patient. In all cases, all potential pre-surgery predictors were set to have no effect on the outcome so that all effects identified by either procedure were false positives.

For each patient we generated a set of predictors based on either the test scores structure or the latent factor scores structure of our data set. Because there were seven independent latent factors extracted from our data by EFA with varimax rotation, the first set of potential pre-surgery predictors consisted of seven independent Gaussian variables with mean zero and unit variation (see the first row of Figure 12). Regarding potential pre-surgery predictors based of the test scores structure of our data set, we opted to generate two distinct sets of such potential predictors, one consisted of twenty-three independent variables (see the second row of Figure 12) while the other consisted of twenty-three covaried variables (see the third row of Figure 12). In both cases, the potential pre-surgery predictors were generated from Gaussian distribution with zero mean and unit variance. Twenty-three independent predictors were generated to test the "best case scenario" whereby data satisfy the assumption of independence of predictors implicit in both the **univariable screening** procedure and the Bayesian Lasso. Twenty-three covaried predictors were generated to test the more realistic scenario whereby there is a non-zero covariance structure among potential pre-surgery predictors derived from single test scores (which unlike the varimax rotated factor analysis results do not invoke statistical independence). In these simulations, we opted

to generate potential predictors via a multivariate Gaussian distribution with zero marginal means, unit marginal variances and a minimal covariance structure whereby predictors representing test scores derived from the same task (e.g., TMT-A and TMT-B) share about 50% of variance while potential predictors representing test scores derived from distinct tasks (e.g., TMT-A and RAVLT-IR) do not share any variance (see Figure 11 for the exact covariance structure used for our simulations).

Statistical models

For each combination of population-level slope b_1 (none, mild and moderate decline) and potential pre-surgery predictor structure (seven independent, twenty-three independent and twenty-three covaried predictors), total of one hundred **univariable screening** procedure and one hundred Bayesian Lasso models were fitted on the same one hundred simulated data sets with null effect of each potential predictor. For the **univariable screening** procedure, first an independent linear mixed model (LMM) with correlated patient-level intercepts and slopes was fitted for each predictor (including the effect of time, predictor and their interaction) and if the p-value of the interaction term between predictor and time showed $p < .2$ (the results were not sensitive to this threshold as the reader can validate by running the code themselves while changing the threshold) the predictor was then entered into a multiple LMM with all such predictors included at the same time; all predictors for which their interaction term with time showed $p < .05$ in this second multiple regression LMM were declared significant and constituted a false positive error. For the Bayesian Lasso, single LMM with correlated patient-level intercepts and slopes including all potential predictors, time and their interactions was fitted and all predictors with probability of direction $>$ 2.5 % (which is equivalent to two-sided $p < .05$) were declared significant and constituted a false positive error.

Results

The results of simulations are presented in Figure 12. It is clear that both factors theorized to alleviate the false positive error rates (i.e., the Bayesian Lasso and dimension reduction of the potential predictors structure) can do so in our data

structure according to these simulations. Moreover, applying the **univariable screening** procedure to our data set seems to incur a high risk of inflated false positive error rates even in the best case scenario.

Figure 1

Factor analyses fit indexes. Density plots of (A) Tucker-Lewis Index (TLI) and (B) upper bound of 90% confidence interval (CI) of the root-mean-square-error approximation for three- to eight-factor solutions of factor analysis of pre-surgery cognitive profile. Density plots are taken over one hundred imputed datasets. Vertical lines represent boundaries of good fit according to TLI (i.e., TLI > 0.9) and RMSEA (i.e., RMSEA < 0.08).

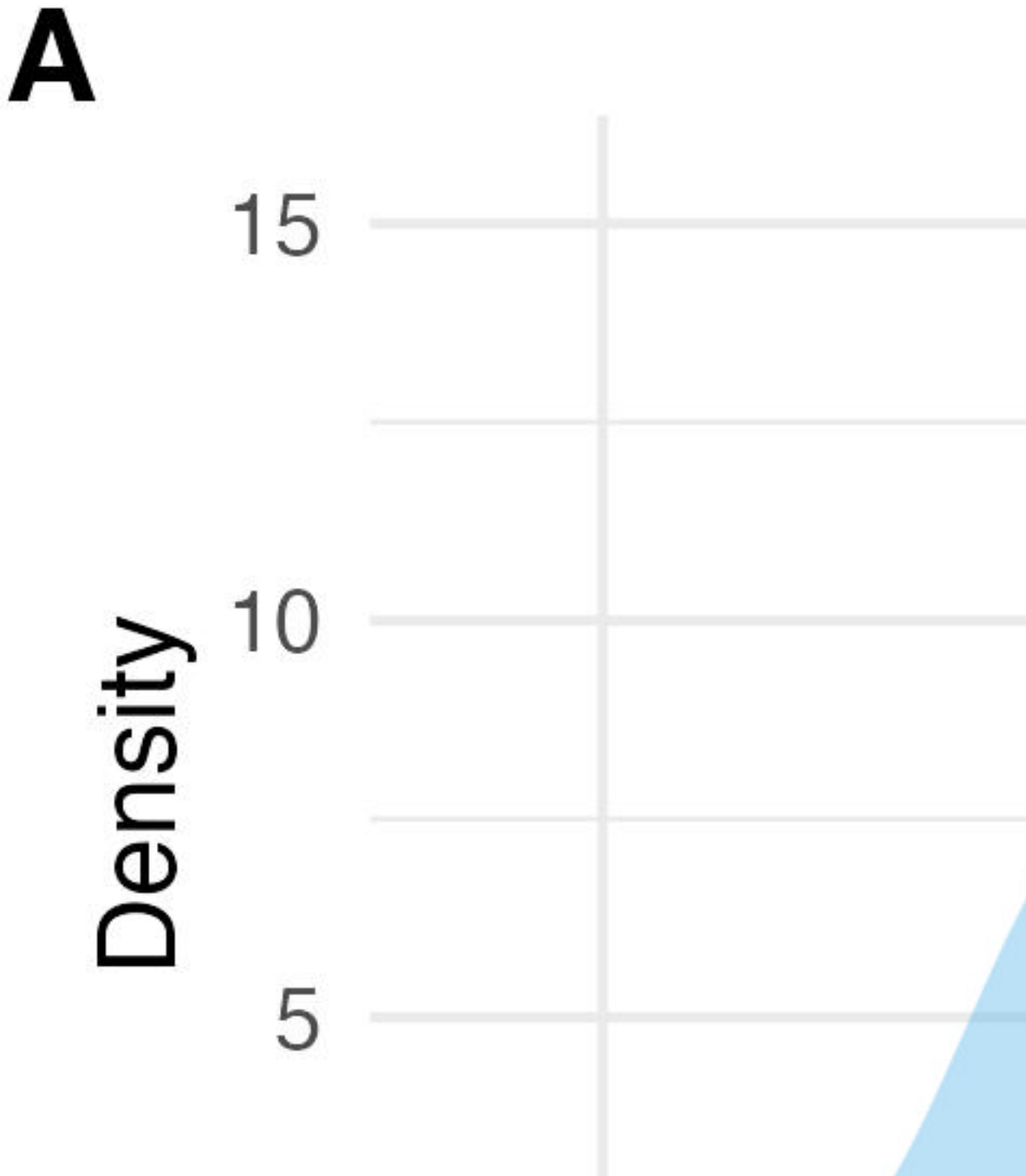

Figure 2

Posterior predictive checks. Posterior predictions of included patients' performance according to the descriptive and predictive generalized linear mixed models (GLMMs) reported in the main text. Lines represent expected (median) performance, shades represent 95% posterior probability intervals (PPIs) of the performance according to the GLMMs, dots represent observed values.

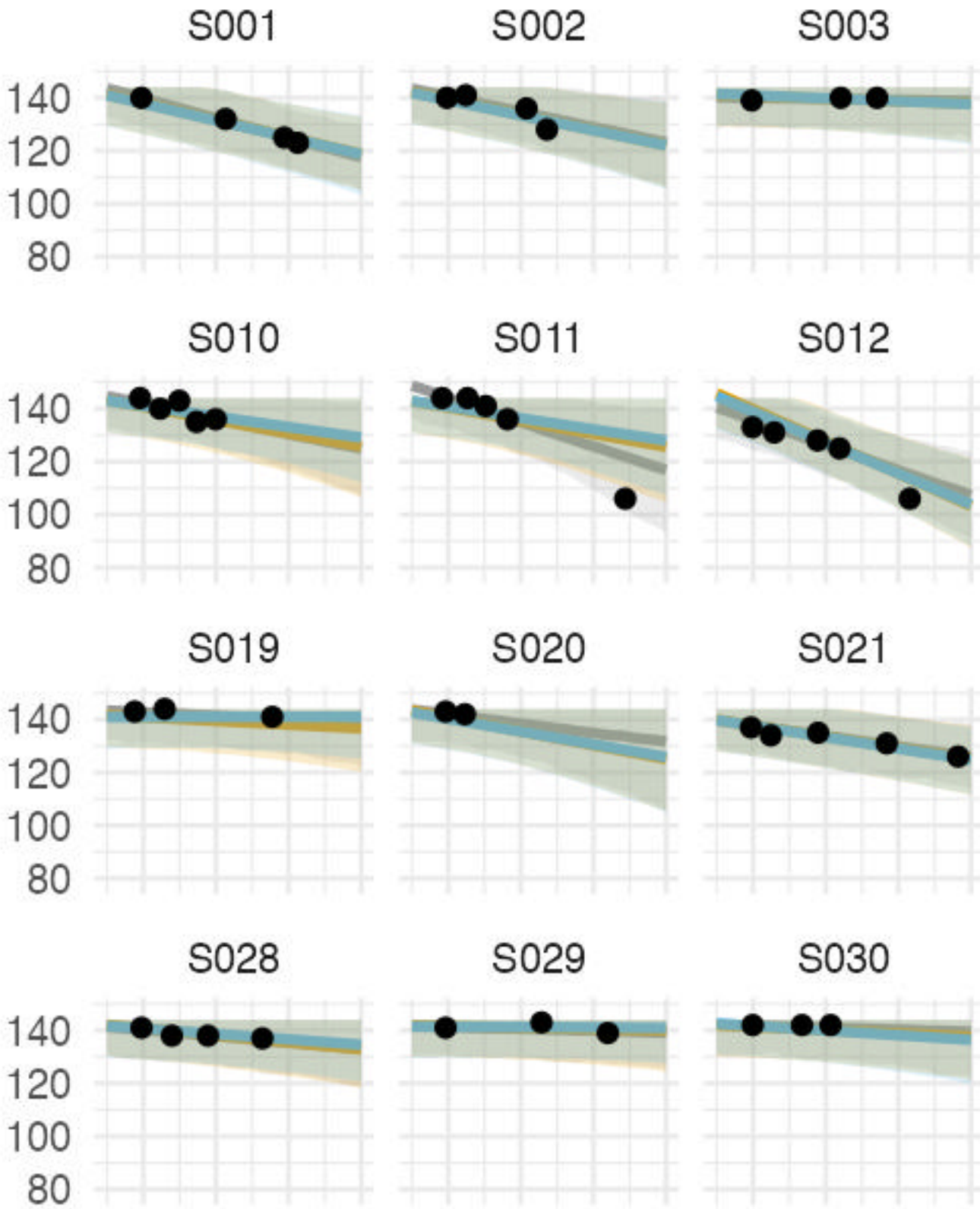

Figure 3

Posteriors medians and 90% equal-tailed intervals (ETIs) of post-minus-pre-surgery differences at monthly post-surgery assessments separately for the sample (dark pink) and population (medium pink) version of our RQ1 estimand. Light pink lines represent model's predictions of raw data differences (change scores).

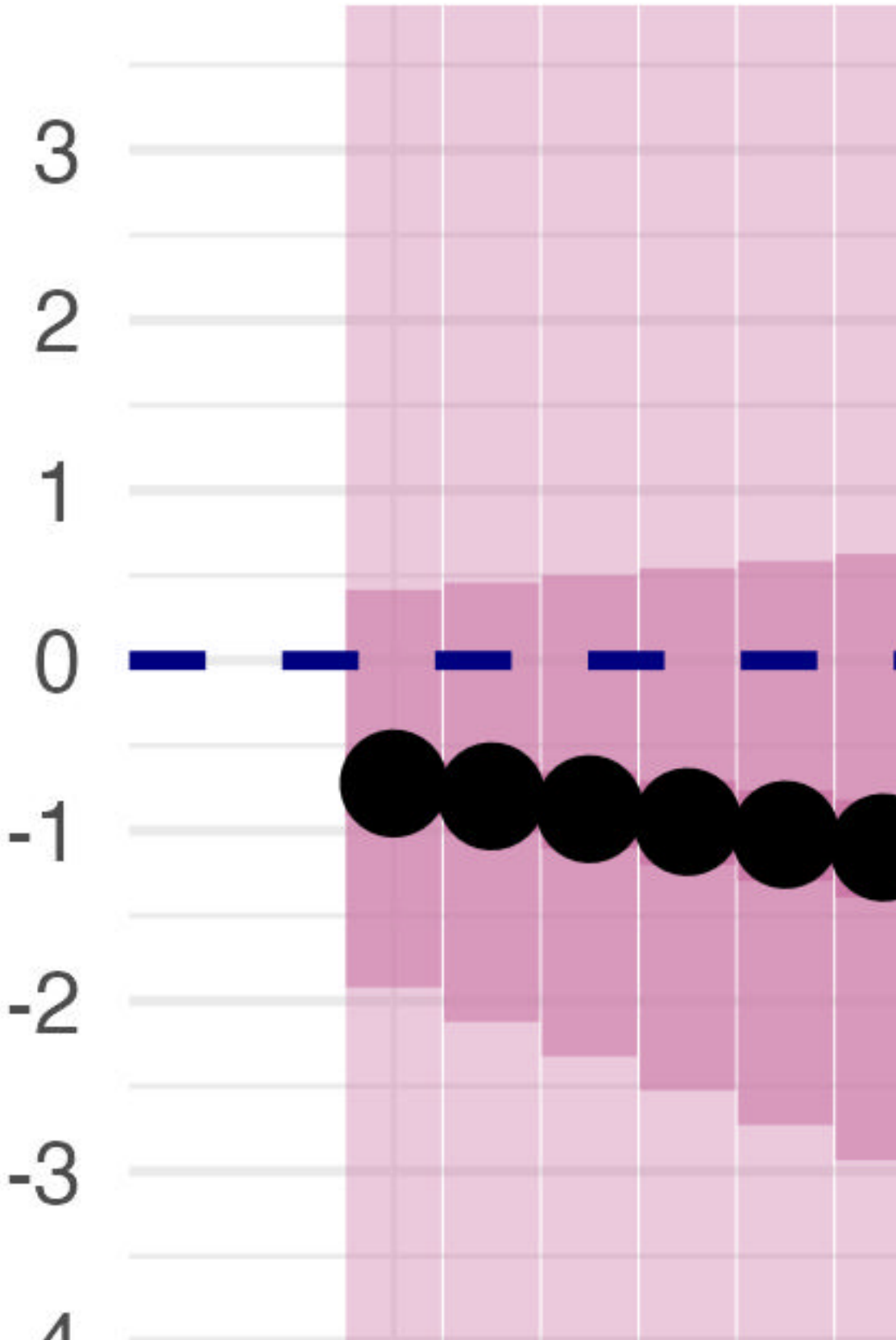

Figure 4

Posterior distributions of patient-level true score variability after accounting for group-level association with time (time only model) or time and pre-surgery cognitive profile (test and factor scores). The plot includes standard deviations of patient-level intercepts (top row), slopes (middle row) and intercept/slope correlations (bottom row).

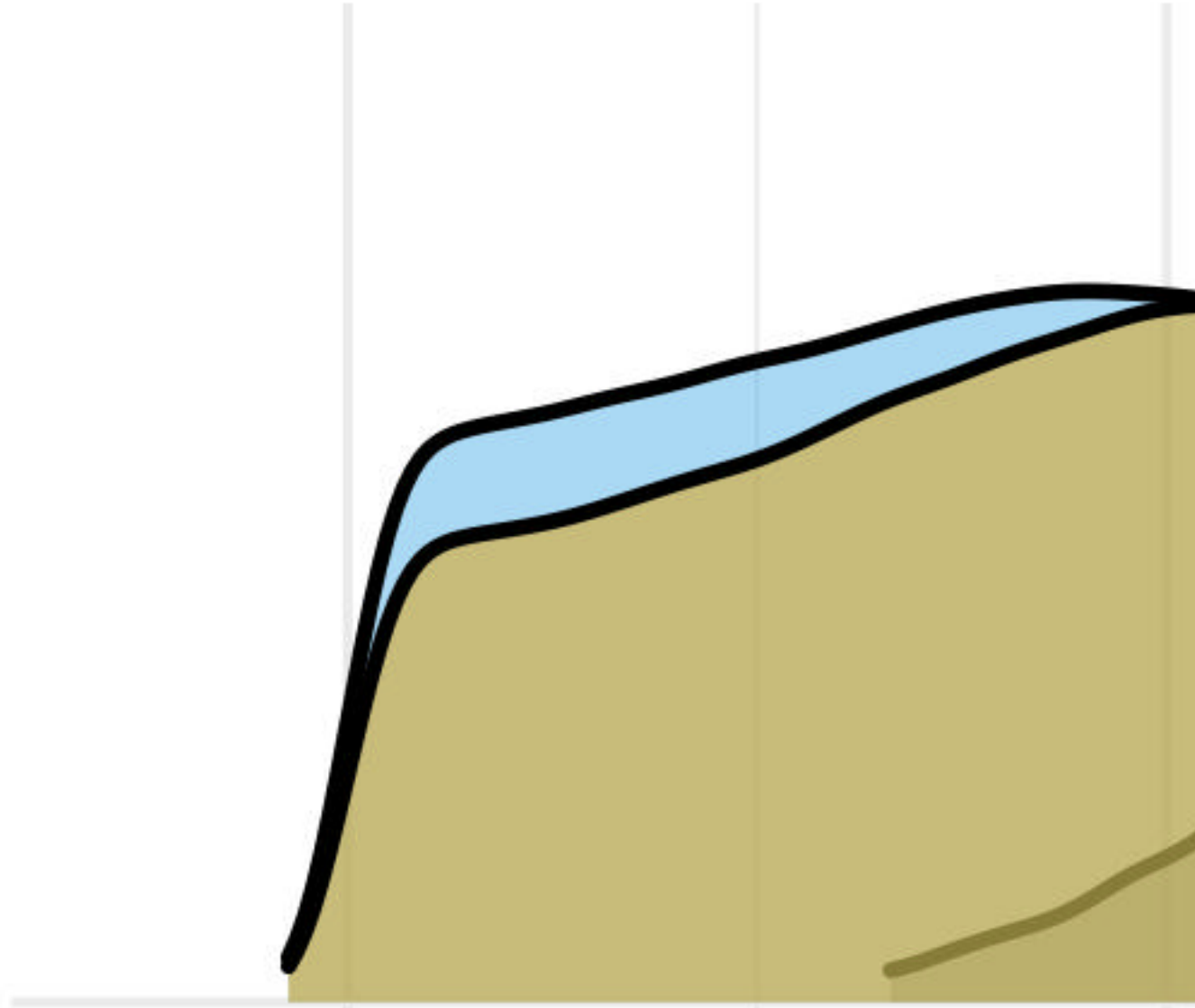

Figure 5

Robustness check (the "test scores" model). Posteriors medians and 95% posterior probability intervals (PPIs) of group-level parameters from the longitudinal generalized linear mixed model predicting post-surgery cognitive decline by pre-surgery cognitive test scores without ("test scores") and with adjustment for covariates ("test scores (with covariates)"). All cognitive predictors were scaled such that negative values mean negative effect of pre-surgery deficit on prediction of longitudinal cognitive trajectory. See main text for acronyms.

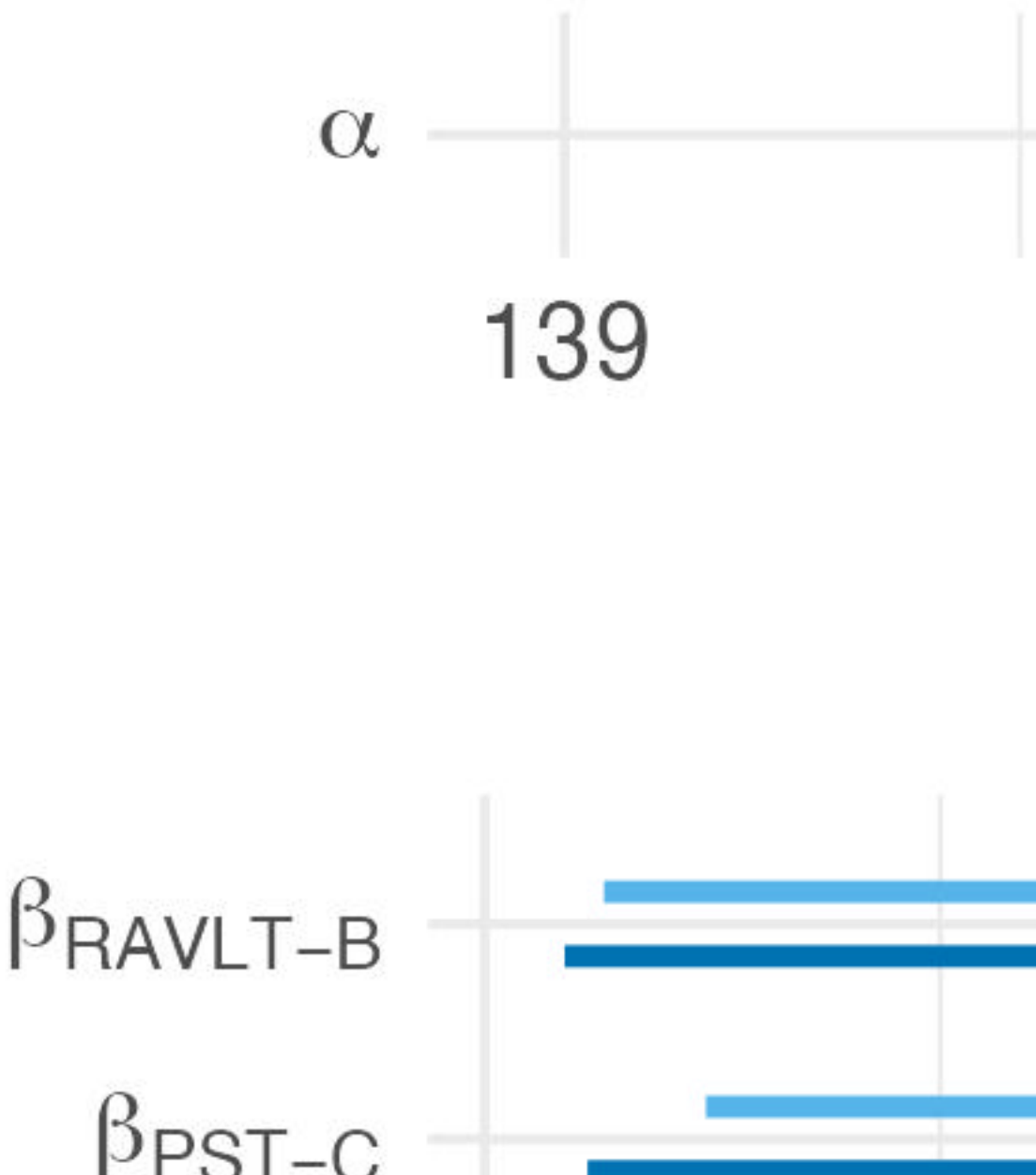

Figure 6

Robustness check (the "factor scores" model). Posteriors medians and 95% posterior probability intervals (PPIs) of group-level parameters from the longitudinal generalized linear mixed model predicting post-surgery cognitive decline by pre-surgery latent cognitive factor scores without ("factor scores") and with adjustment for covariates ("factor scores (with covariates)"). All cognitive predictors were scaled such that negative values mean negative effect of pre-surgery deficit on prediction of longitudinal cognitive trajectory. See main text for acronyms.

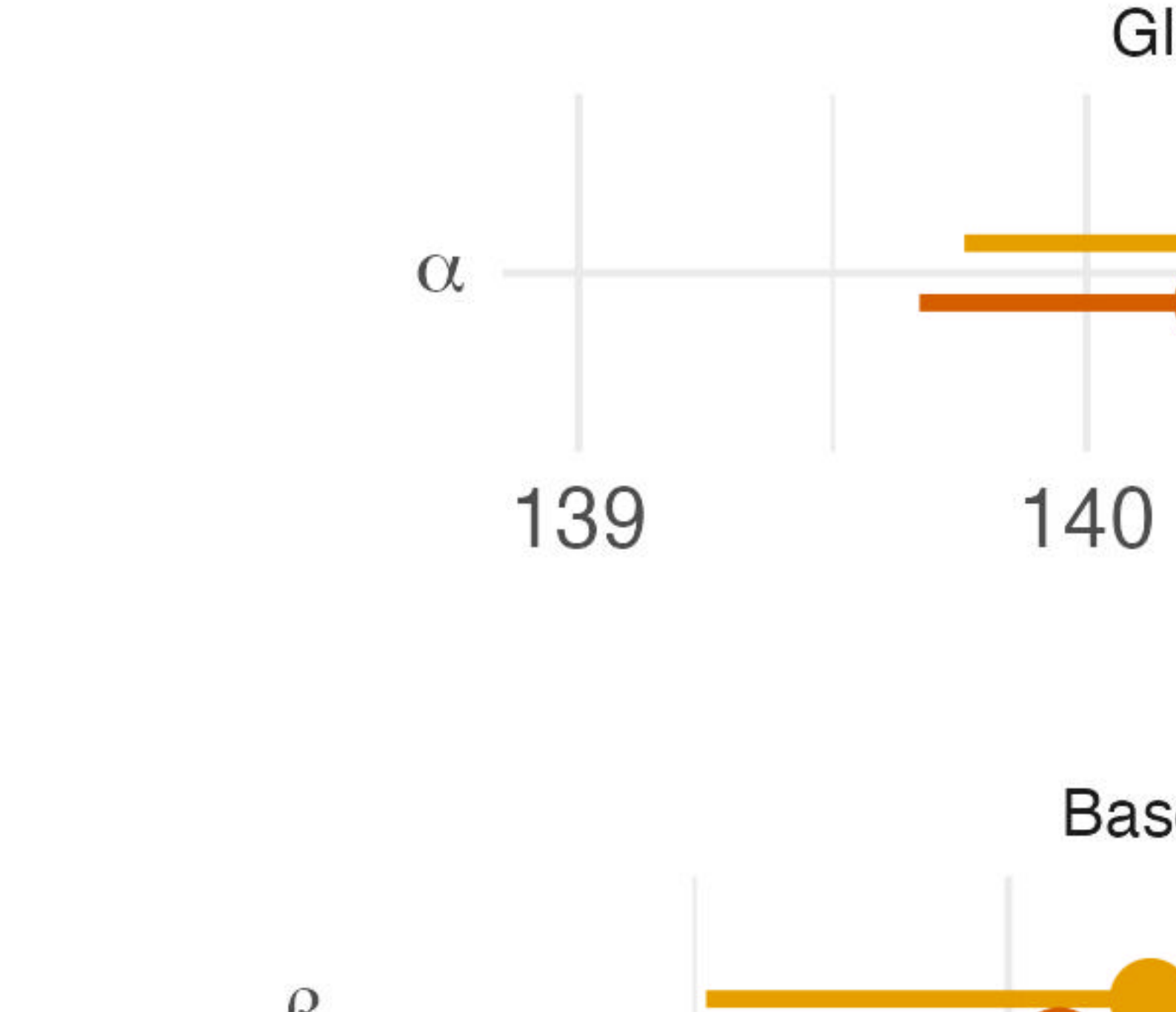

Figure 7

Position of electrodes as estimated by Lead-DBS.

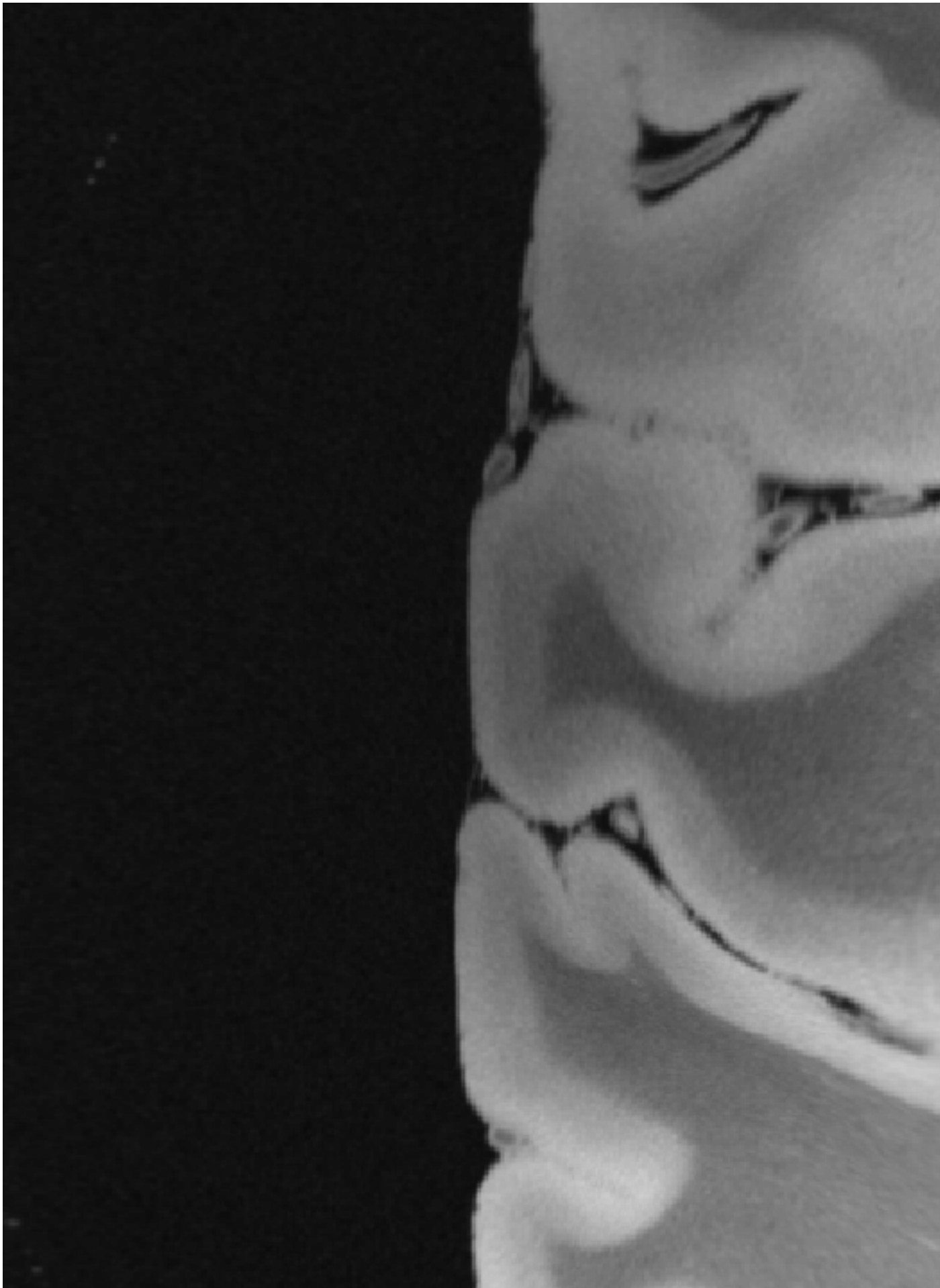

Figure 8

Between-group comparisons of included vs excluded patients. Bayesian p-values were calculated as half of the probability that the difference parameter (described by its posterior distribution) is strictly positive or negative. Lower values imply higher probability of difference between groups. Conventional $< .05$ value is marked by the red dashed line. The left column relates to mean differences while the right column relates to differences in standard deviations.

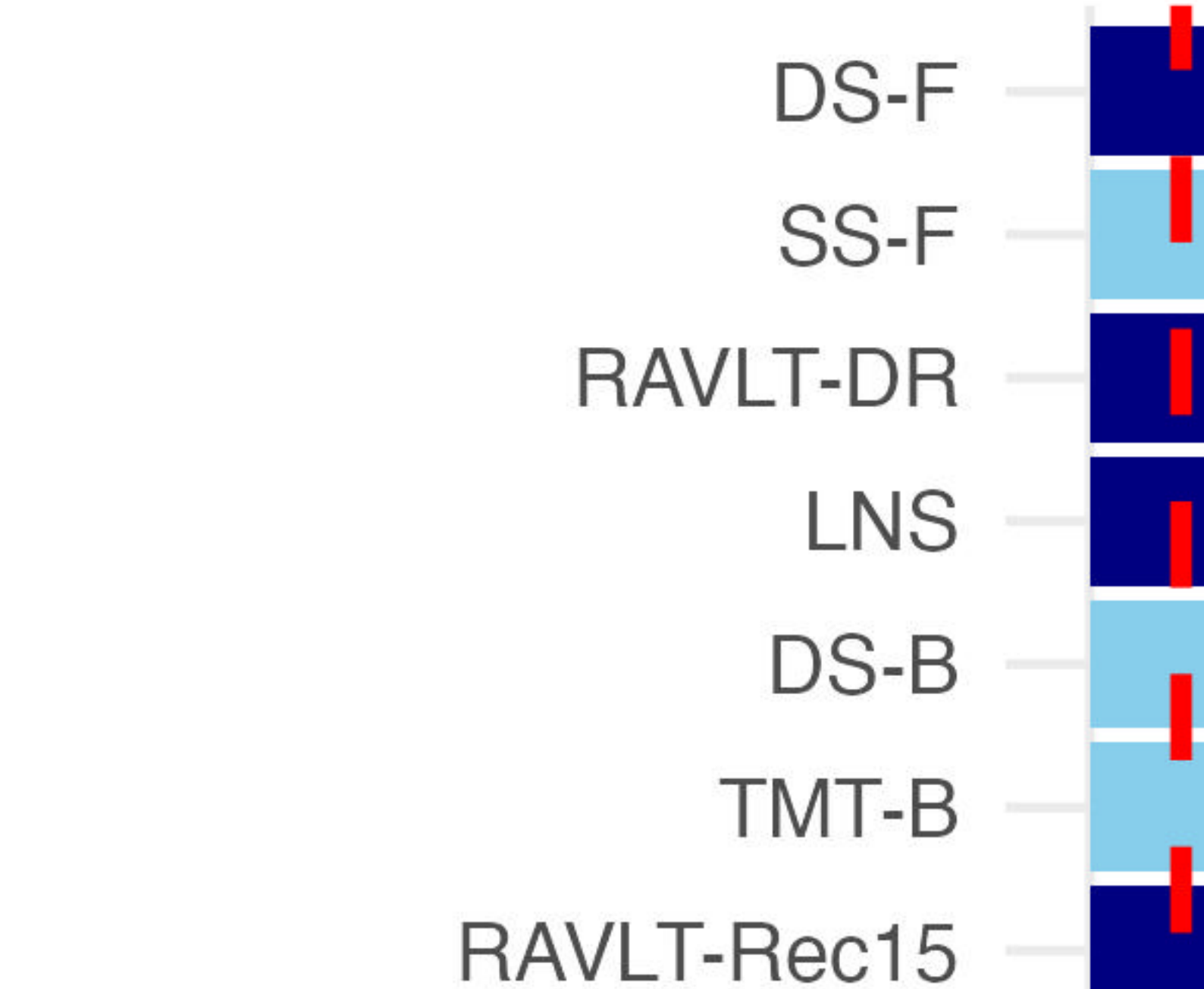

Figure 9

Full posterior distributions of interaction terms of the model predicting post-surgery cognitive decline by proportion of subthalamic nucleus (STN) components that is being stimulated. All posteriors were scaled such that effects associated with each predictor represent comparisons of expected yearly cognitive decline between patients differing by ten percentage points in overlap of affected proportion of the relevant STN component. Acronyms are explained in Table S6.

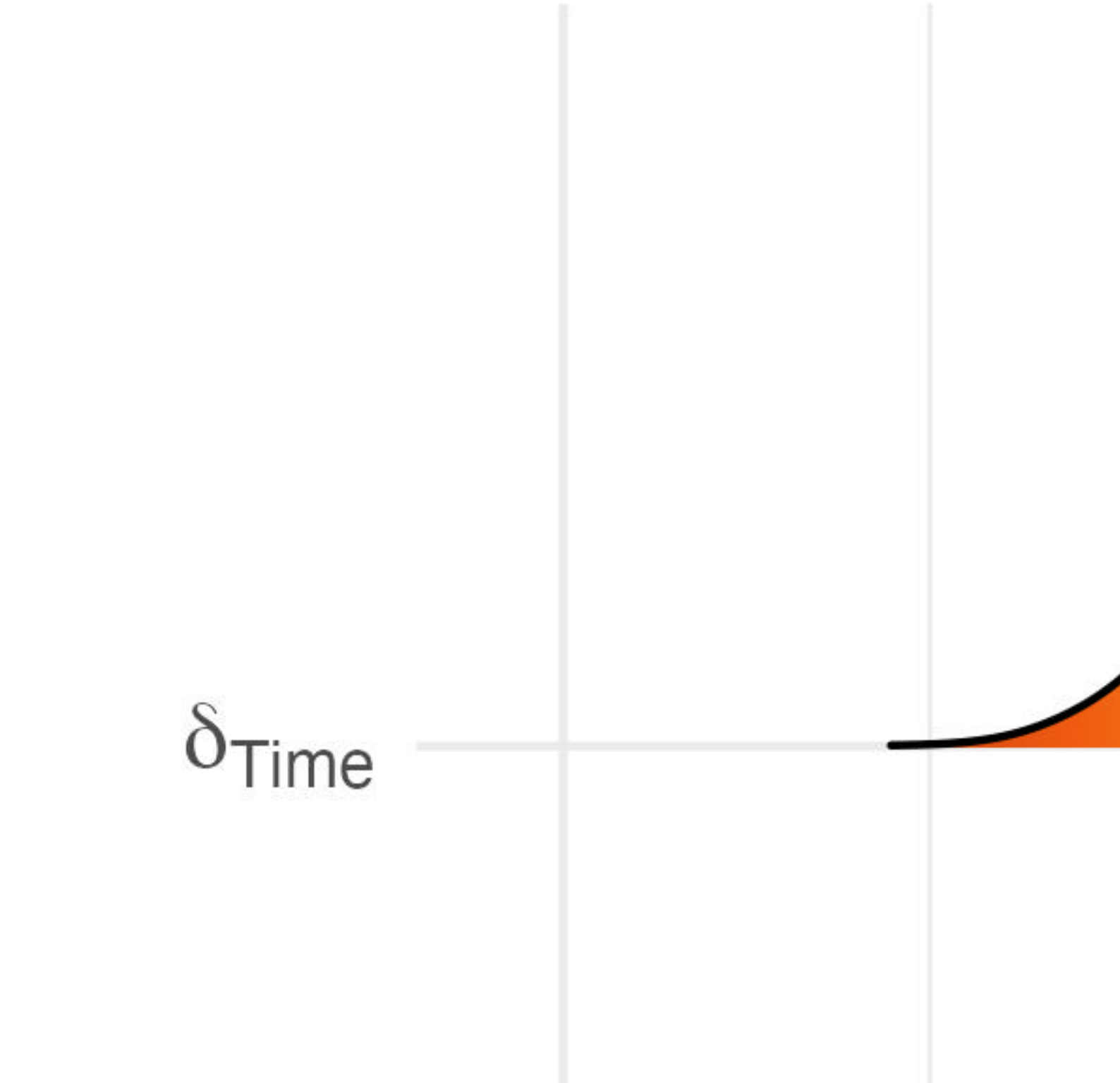

Figure 10

Collinearity metrics of group-level parameters of the model predicting cognitive decline by the proportion of subthalamic nucleus components volume that is affected by stimulation.

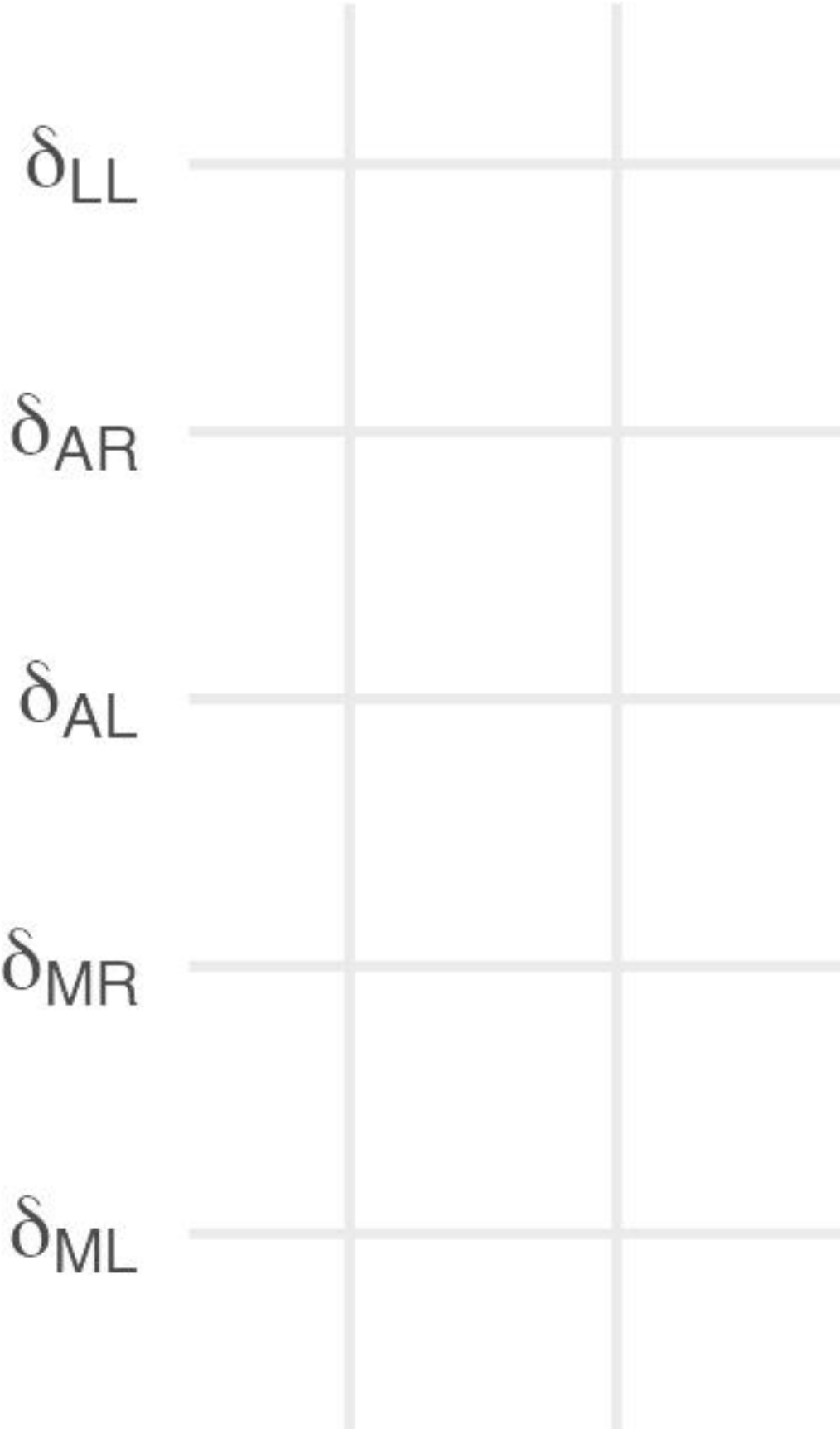

Figure 11

Covariance matrix of the covaried test scores predictor structure. The figure represents correlations used for generation of covaried predictors in the covaried test scores data-generating process. Orange clusters represent high correlations among State-Trait Anxiety Inventory, Tail Making Test, Digit Span, Spatial Span, Stroop task, verbal fluency, Rey Auditory Verbal Learning Test and Family Pictures test respectively. The single non-correlated cells represent the Tower of London task and Similarities task.

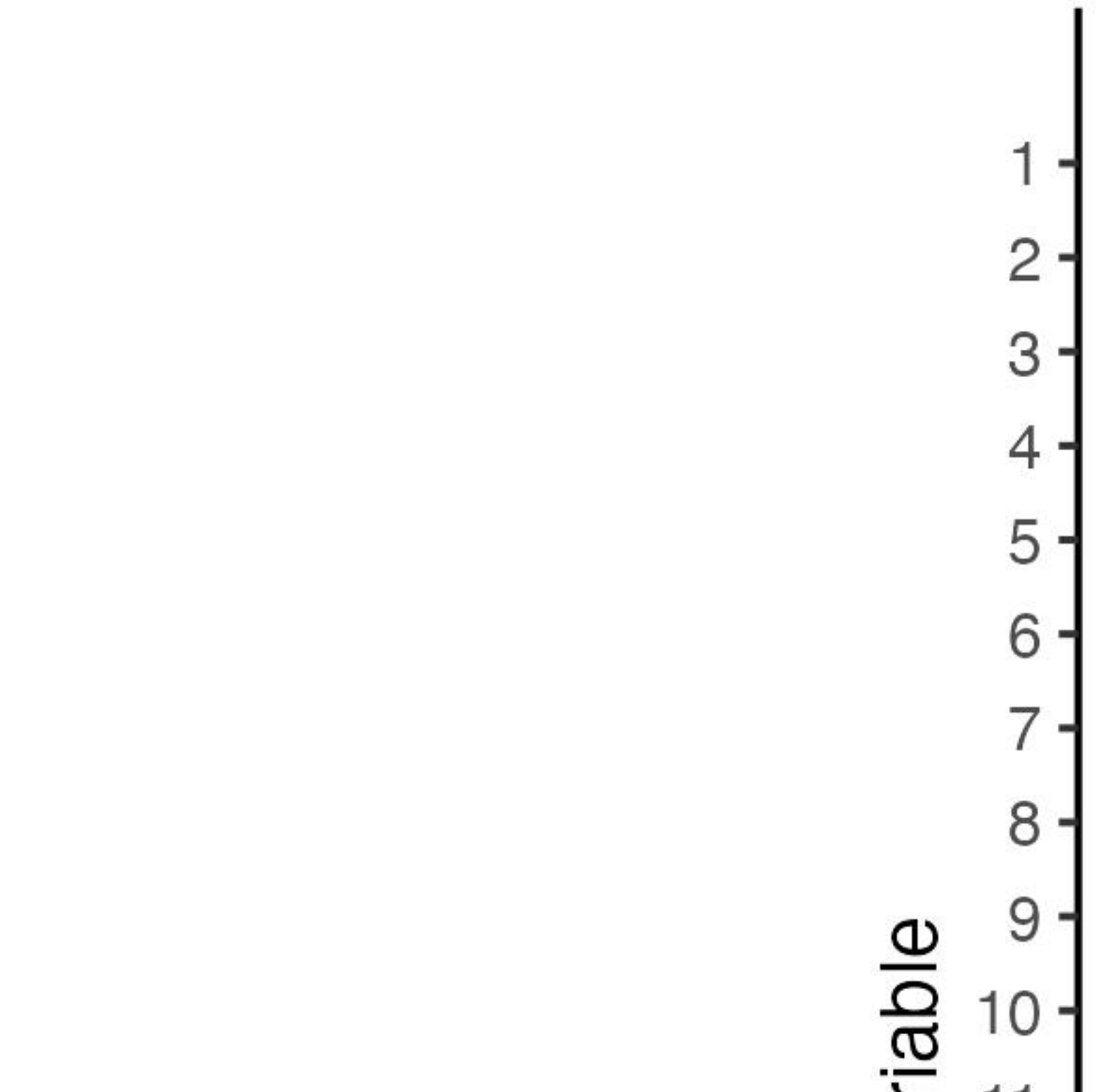

Figure 12

Simulation results. The figure presents number of false positives per one hundred simulations dependent on (i) method used (colour), (ii) assumed average annual post-surgery decline (columns) and (iii) potential pre-surgery predictor structure (rows).