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], 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, Department of Neurology and Centre of Clinical Neuroscience, First Faculty of Medicine and General University Hospital in Prague, Charles University, Czech Republic, 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 manuscript *"Preoperative Cognitive Profile Predictive of Cognitive Decline after Subthalamic Deep Brain Stimulation in Parkinson's Disease"* including further presentation of the 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_longCOG. 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 the 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 (minres) solution. This procedure was repeated for three up to eight factor solutions.

Supplementary presentation of results

Supplementary EFA results are presented in Table S1 and Figure S1 (see below). Table S1 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 out of four 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 S1 which depicts density plots of TLI and upper 90% confidence interval boundary 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 S1

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

Model	TLI	RMSEA	RMSEA 90% CI (upper bound)	Total variance accounted for	upper bound RMSEA < 0.08 (%)	TLI > 0.90 (%)
3-factor	0.68 (0.03)	0.09 (0.00)	0.11 (0.00)	0.38 (0.01)	0	0
4-factor	0.81 (0.03)	0.07 (0.01)	0.09 (0.00)	0.44 (0.01)	6	0
5-factor	0.87 (0.03)	0.06 (0.01)	0.08 (0.01)	0.48 (0.01)	68	16
6-factor	0.92 (0.03)	0.04 (0.01)	0.07 (0.01)	0.52 (0.01)	96	74
7-factor	0.96 (0.03)	0.03 (0.01)	0.06 (0.01)	0.55 (0.01)	99	97
8-factor	0.99 (0.03)	0.02 (0.01)	0.05 (0.01)	0.58 (0.01)	100	100

Values represent mean (SD) or percentages if indicated in brackets.
TLI Tucker-Lewis Index. RMSEA root-mean-square-error approximation. CI confidence interval

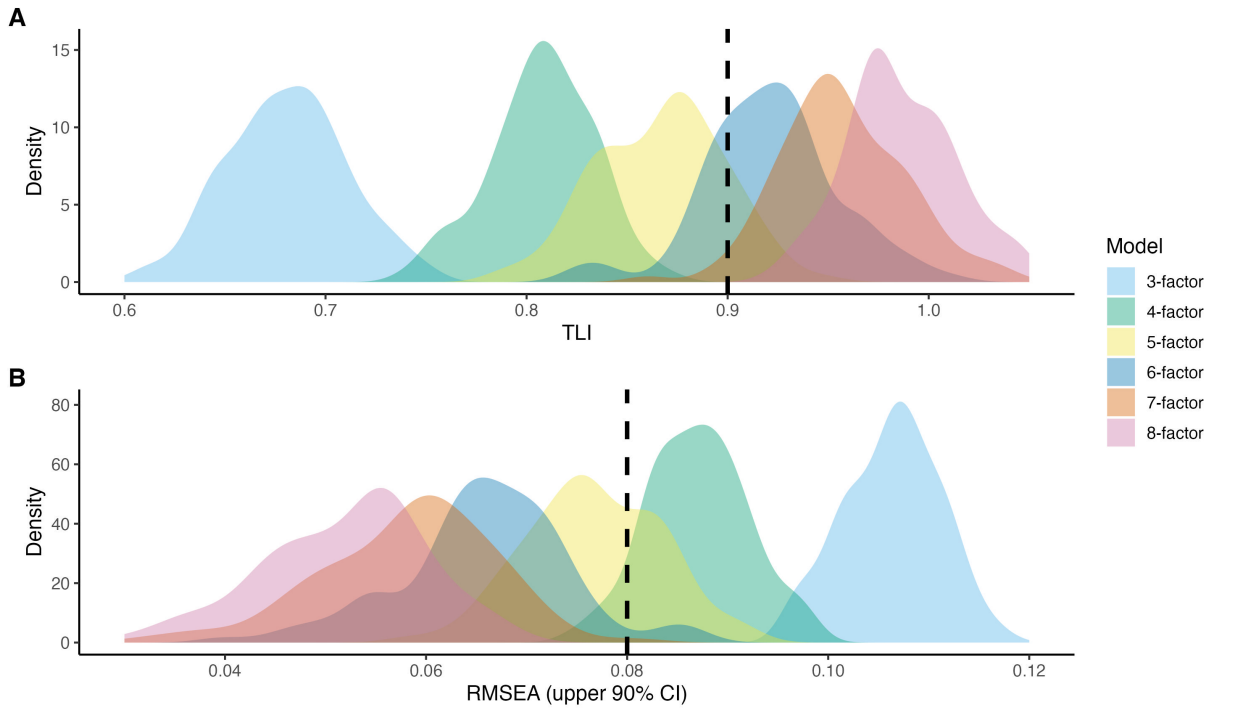

Figure S1

Factor analyses fit indexes. Density plots of (A) Tucker-Lewis Index (TLI) and (B) upper boundary 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).

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 in 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 S3, Figure S4, Table S2, Table S3 as well as Figure 4 in the main text) thus represent an effect of a (relative) pre-surgery deficit in a 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 a pre-surgery cognitive deficit unfavorably affects the outcome and vice versa for positive parameter values.

Posterior predictive check

To validate the in-sample fit of our predictive models, we computed models' "predictions" for each included patient and compared these predictions to observed values (see Figure S2). 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, the model is 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 S2). Furthermore, while the "test scores" and the "factor scores" model provide similar posterior predictions for our patients, the "test scores" model was evidently more influenced by outlying values to a small degree (for instance patients S023, S107 or S124).

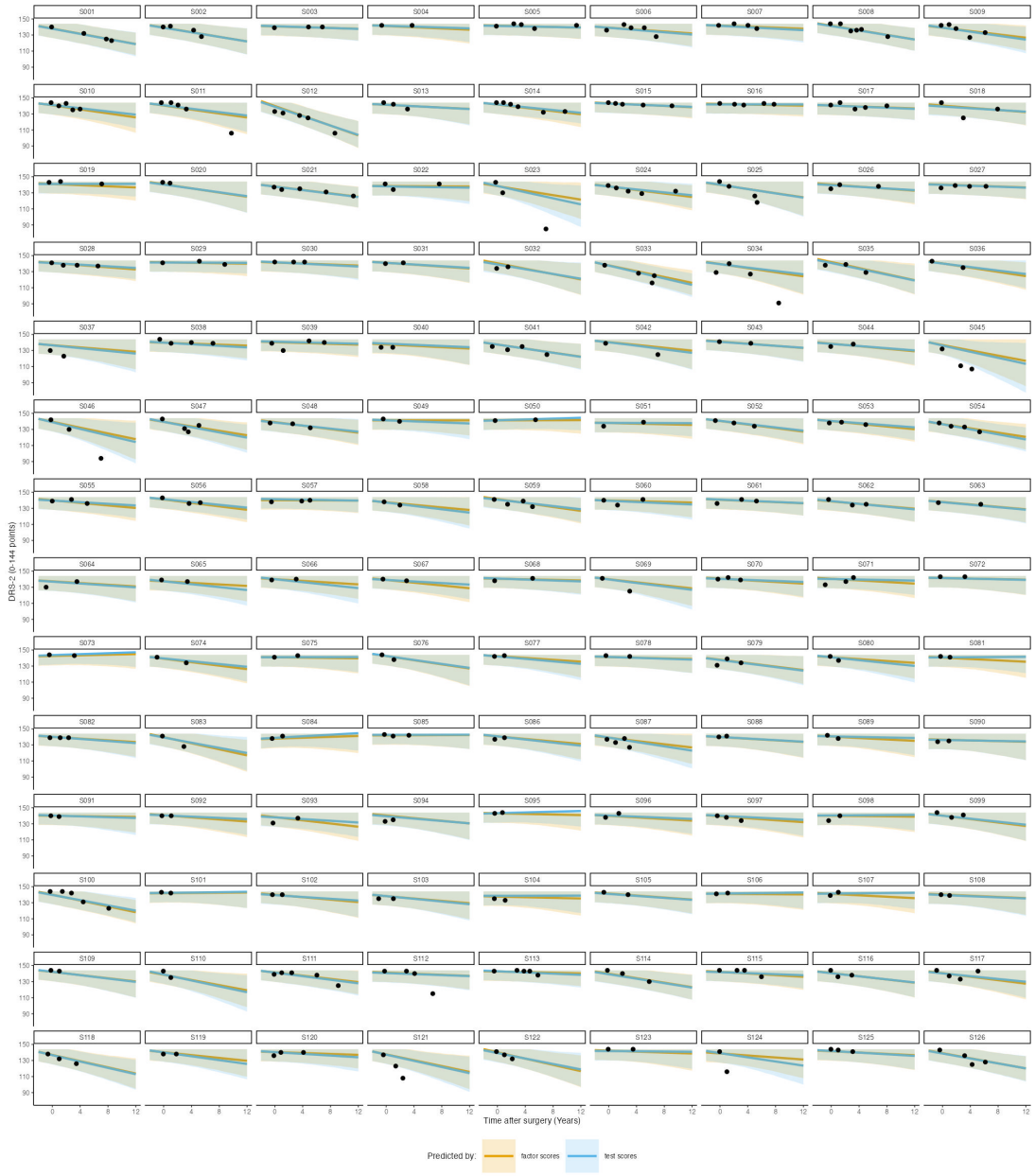

Figure S2

Posterior predictive checks. Posterior predictions of included patients' performance according to the 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.

Supplementary presentation of results

In Table S2 we present numerical summary of group-level posterior parameters of the "test scores" model while in Table S3 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 S2

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

Parameter	*b*	95% PPI	Pr(*b*<0)
Global intercept (α)			
Intercept	140.17	[139.58, 140.75]	0.000
Baseline correlates (β)			
TMT-A	0.00	[-0.35, 0.37]	0.486
TMT-B	-0.26	[-0.86, 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	[-0.38, 0.41]	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

Table S2

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

Parameter	*b*	95% PPI	Pr(*b*<0)
RAVLT-B	-0.35	[-0.96, 0.10]	0.936
RAVLT-DR	0.06	[-0.28, 0.54]	0.330
RAVLT-Rec50	-0.01	[-0.40, 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	[-0.35, 0.38]	0.472
Time-dependent effects (δ)			
Time	-0.72	[-0.98, -0.47]	1.000
TMT-A × Time	-0.08	[-0.32, 0.11]	0.810
TMT-B × Time	-0.15	[-0.45, 0.09]	0.887
DS-F × Time	0.12	[-0.08, 0.35]	0.130
DS-B × Time	0.07	[-0.14, 0.32]	0.242
LNS × Time	0.07	[-0.15, 0.33]	0.255
SS-F × Time	0.23	[-0.06, 0.57]	0.057
SS-B × Time	-0.11	[-0.39, 0.11]	0.838
TOL × Time	-0.05	[-0.27, 0.14]	0.696
PST-D × Time	-0.01	[-0.27, 0.22]	0.561
PST-W × Time	-0.15	[-0.44, 0.08]	0.896
PST-C × Time	-0.09	[-0.35, 0.12]	0.811
COWAT × Time	-0.14	[-0.36, 0.05]	0.922
CFT × Time	-0.02	[-0.24, 0.20]	0.581
Sim. × Time	0.08	[-0.13, 0.34]	0.227
RAVLT-IR × Time	0.00	[-0.23, 0.24]	0.478
RAVLT-B × Time	0.02	[-0.17, 0.24]	0.392
RAVLT-DR × Time	0.07	[-0.13, 0.31]	0.228

Table S2

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

Parameter	*b*	95% PPI	Pr(*b*<0)
RAVLT-Rec50 × Time	-0.03	[-0.27, 0.17]	0.632
RAVLT-Rec15 × Time	-0.00	[-0.22, 0.21]	0.521
FP-IR × Time	-0.03	[-0.34, 0.26]	0.594
FP-DR × Time	-0.05	[-0.39, 0.22]	0.684
STAI-X1 × Time	-0.00	[-0.20, 0.18]	0.521
STAI-X2 × 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 media); 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); ×: 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 S3

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

Parameter	*b*	95% PPI	Pr(*b*<0)
Global intercept (α)			
Intercept	140.25	[139.68, 140.84]	0.000
Baseline correlates (β)			
EF/Att.	-0.19	[-0.78, 0.28]	0.790
EM	-0.17	[-0.71, 0.25]	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 (δ)			
Time	-0.75	[-0.99, -0.51]	1.000
EF/Att. × Time	-0.39	[-0.63, -0.15]	0.999
EM × Time	-0.00	[-0.22, 0.22]	0.510
VWM × Time	0.17	[-0.09, 0.45]	0.099
VM × Time	-0.17	[-0.44, 0.10]	0.888
SS × Time	-0.14	[-0.47, 0.18]	0.779
An. × Time	-0.00	[-0.21, 0.21]	0.504
SWM × 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 media); 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); ×: 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). Stan code for each of these models with the exact specification is available at https://github.com/josefmana/dbs_longCOG.git and is equivalent to the models reported in the main text. For this reason, we do not present their mathematical definitions here in sake of brevity.

Side-to-side comparison of models' group-level parameters' posterior summaries is presented in Figure S3 for the "test scores" model and in Figure S4 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.

Estimation of false positive rates

To test our assumption that the classical two-step 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 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

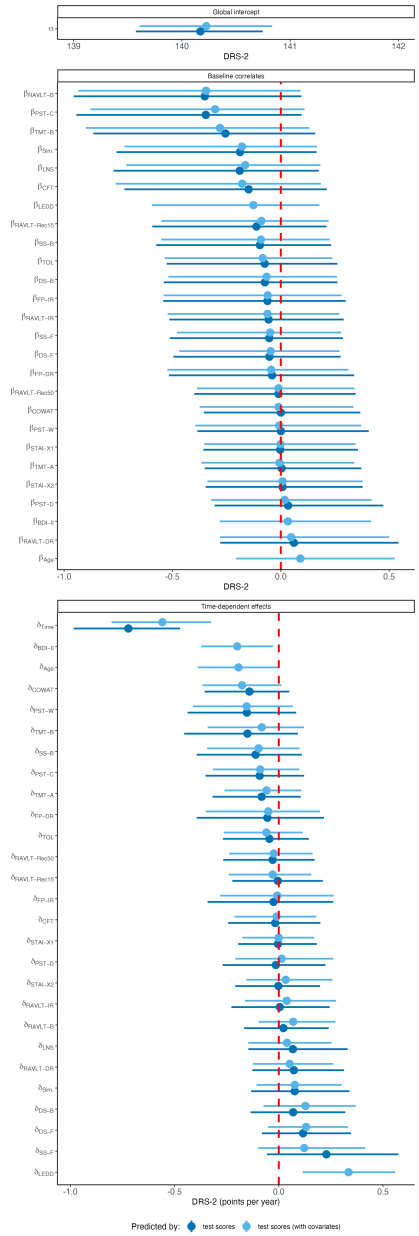

Figure S3

Posteriors medians and 95% posterior probability intervals (PPIs) of group-level effects 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 longitudinal cognitive trajectory. See main text for acronyms.

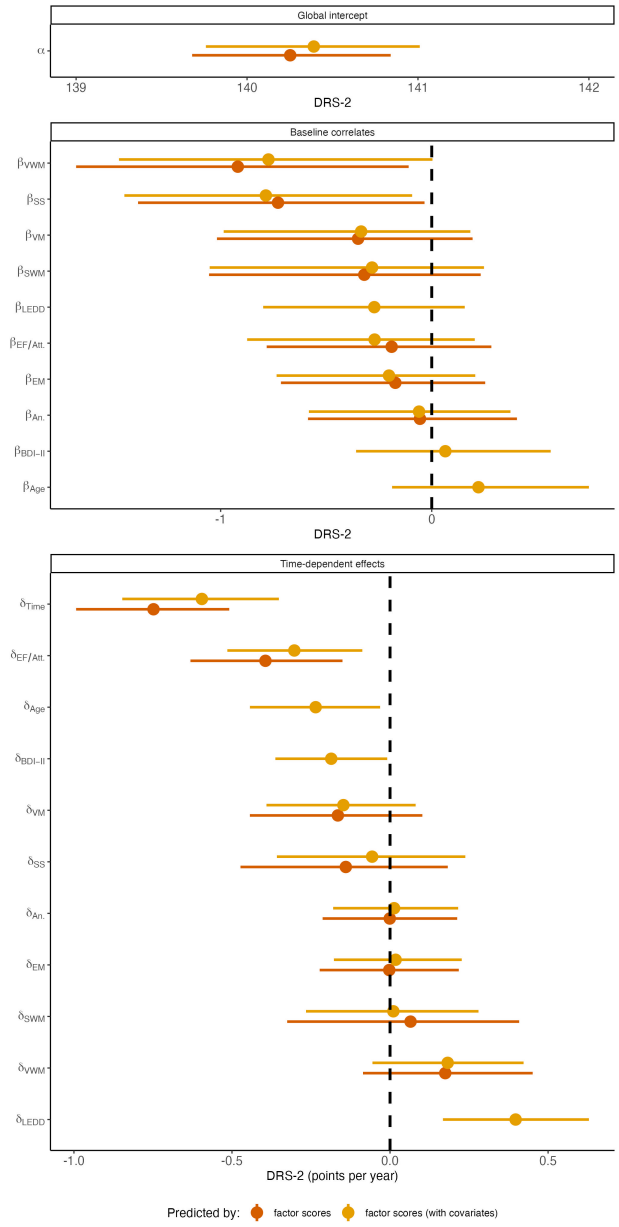

Figure 4

Robustness check (the "factor scores" model). Posteriors medians and 95% posterior probability intervals (PPIs) of group-level effects 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 longitudinal cognitive trajectory. See main text for acronyms.

data sets used to produce our results as well as generating functions are available at https://github.com/josefmana/dbs_longCOG.git. The reader should navigate to the "dbs_longCOG_optbias.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 S6). 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 the two-step procedure or the Bayesian Lasso 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 S6). 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 S6) while the other consisted twenty-three covaried variables (see the third row of Figure S6). 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 two-step 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 the 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 S5 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 two-step 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 two-step procedure, first an independent linear mixed model (LMM) with correlated patient-level intercepts and slopes was fitted for each predictor

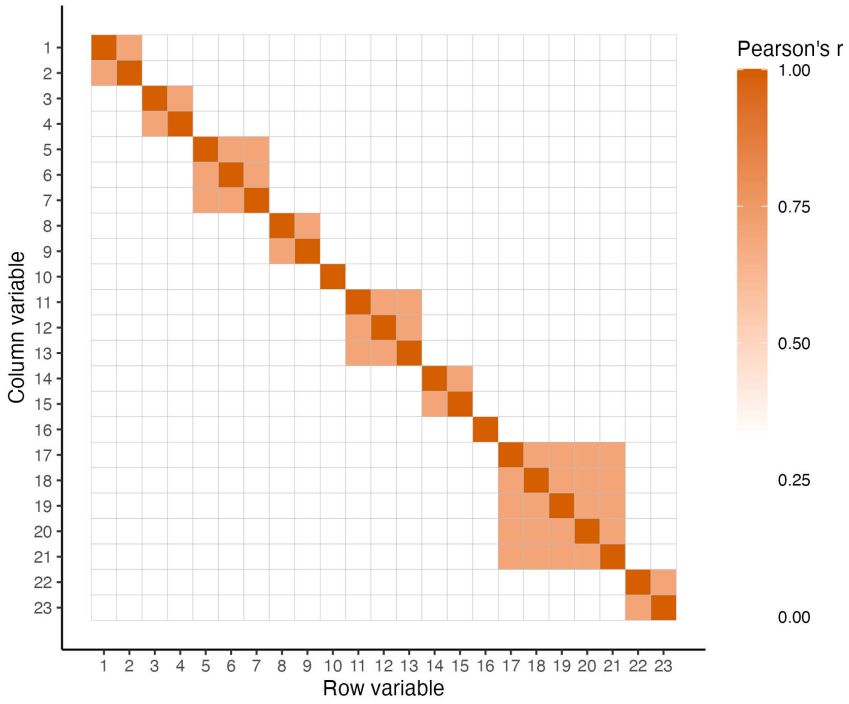

Figure S5

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. The 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.

(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, a 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.

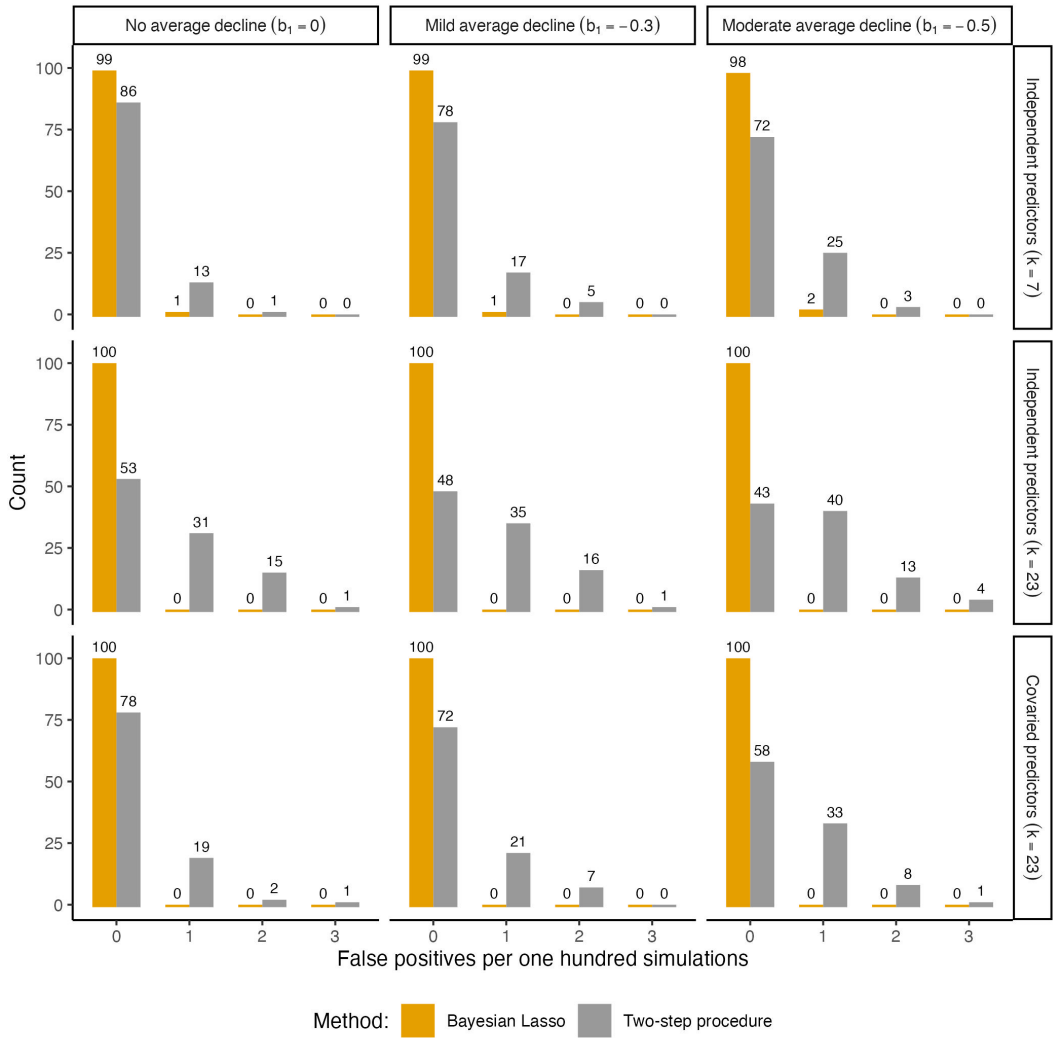

Figure 6

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

Results

The results of simulations are presented in Figure S6. 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 two-step procedure to our data set seems to incur a high risk of inflated false positive error rates even in the best case scenario.