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], Andrej Lasica[1], Anna Smidova[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

Josef Mana https://orcid.org/0000-0002-7817-3978

Ondrej Bezdicek https://orcid.org/0000-0002-5108-0181

Robert Jech https://orcid.org/0000-0002-9732-8947

Author roles were classified using the Contributor Role Taxonomy (CRediT; https://credit.niso.org/) as follows: *Josef Mana*: Conceptualization, Data curation, Investigation, Formal analysis, Methodology, Software, Visualization, Writing - original draft, Writing - review & editing; *Ondrej Bezdicek*: Conceptualization, Data curation, Investigation, Methodology, Supervision, Writing - original draft, Writing - review & editing; *Filip Ruzicka*: Investigation, Writing - review & editing; *Andrej Lasica*: Investigation, Data curation, Formal analysis, Writing - review & editing; *Anna Smidova*: Investigation; *Olga Klempirova*: Investigation; *Tomas Nikolai*: Investigation, Writing - review & editing; *Tereza Uhrova*: Investigation; *Evzen Ruzicka*: Conceptualization, Funding acquisition, Investigation, Writing - review & editing; *Dusan Urgosik*: Investigation; *Robert Jech*: Conceptualization, Data curation, Funding acquisition, Investigation, Resources, Supervision, Writing - review & editing

Correspondence concerning this article should be addressed to Ondrej Bezdicek, Email: ondrej.bezdicek@gmail.com

Abstract

Cognitive decline represents a severe non-motor symptom of Parkinson's disease (PD) that can significantly reduce benefits of subthalamic deep brain stimulation (STN DBS). Here, we aimed to **describe expected post-surgery cognitive decline and** identify pre-surgery cognitive profile associated with faster post-surgery cognitive decline in STN DBS treated PD patients. A retrospective observational study of 126 PD patients treated by STN DBS combined with oral dopaminergic therapy followed for 3.54 years on average (SD = 2.32) with repeated assessments of cognition was conducted. Pre-surgery cognitive profile was obtained via a comprehensive neuropsychological examination. Data were analyzed using exploratory factor analysis for pre-surgery cognitive profile extraction and Bayesian generalized linear mixed models for description of the longitudinal cognitive outcome. Overall, we observed a mild annual cognitive decline of 0.90 points from a total of 144 points in the Mattis Dementia Rating Scale (95% posterior probability interval (PPI) [-1.19, -0.62]). Pre-surgery executive deficit predicted the rate of post-surgery cognitive decline (b = -0.39, 95% PPI [-0.63, -0.15]). The predictive utility of pre-surgery executive deficit resulted from summing small effects of several single test scores. **Exploratory analysis of electrode localisation did not yield any statistically clear results. Overall, our data and models imply mild average annual post-surgery cognitive decline in PD patients treated with STN DBS with high inter-individual variability. Nonetheless, patients with worse long-term cognitive prognosis can be reliably identified via pre-surgery examination of executive functions.**

Keywords: Parkinson's disease, deep brain stimulation, cognition, longitudinal, hierarchical modelling

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

Introduction

Bilateral subthalamic nucleus (STN) deep brain stimulation (DBS) is an advanced symptomatic treatment of Parkinson's disease (PD) that can successfully reduce motor symptoms and improve patients' quality of life (Armstrong & Okun, 2020; Bratsos et al., 2018). On the other hand, prior research revealed considerable heterogeneity in cognitive outcomes after STN DBS with a small to moderate post-surgery decline in verbal fluency and equivocal results for other cognitive domains (Combs et al., 2015; Mehanna et al., 2017; Parsons et al., 2006). The ability to predict which patients are likely to develop post-surgery cognitive decline can thus prove useful for patient selection and for guiding post-surgery patient monitoring. In this article, we aim to describe **annual rate of long-term post-surgery cognitive decline after STN DBS in PD as well as** pre-surgery cognitive profile extractable from clinically available neuropsychological evaluation that **predicts faster post-surgery decline**.

The majority of prior studies describing longitudinal post-surgery cognitive decline employed pre-surgery/post-surgery design with change scores as their dependent variable (Gruber et al., 2019; Kim et al., 2014) or estimated dementia-free survival time (Barbosa et al., 2024; Bove et al., 2020; Kishore et al., 2019). When conversion to dementia is the outcome, studies do not estimate annual rate of post-surgery decline but the time it takes to reach clinically salient cognitive deficit. On the other hand, change scores have a drawback of confounding true change with measurement error (Singer & Willett, 2003). Furthermore, the focus on change scores allows researchers to estimate group-level post-surgery changes describing their sample but ignores patient-level variability which is necessary to generalize findings beyond the sample (Yarkoni, 2020). In this study, we leverage the fact that our data set includes three or more observations in large enough number of patients to estimate both group-level post-surgery cognitive

decline to describe our sample as well as patient-level variability to provide predictions for other similar samples.

With regards to predicting post-surgery cognitive decline from patients' pre-surgery cognitive profile, studies can be broadly divided to two groups, randomized controlled trials (RCT**s**) and long-term observational studies. In a typical RCT, patients are randomized to treatment and placebo groups and outcomes are compared in a full factorial design (**representing the estimand of interest** as interactions between group and time of assessment) (Schüpbach et al., 2007). Courtesy of their experimental control, RCTs allow for causal inference and are well suited for providing guidelines for patient selection. However, even though RCTs are regarded as a gold standard for causal inference, it is ethically unacceptable to deny DBS treatment for PD patients for longer time intervals than necessary. Long-term (i.e., more than three years after surgery) outcomes can thus be best described by observational studies. While observational studies usually do not allow for causal inference and are not well suited for guiding patient selection due to a lack of proper control group and resulting collider bias (Cinelli et al., 2022), they are well suited for description of patients' long-term outcomes. Longitudinal observational studies can serve as a basis for selecting high-risk STN DBS treated patients that would benefit from increased monitoring.

Previous longitudinal observational studies reported that PD patients treated with STN DBS showing pre-surgery deficit in executive functions **or poorer memory** are at risk of faster post-surgery cognitive decline or developing dementia (Bove et al., 2020; Gruber et al., 2019; Jahanshahi et al., 2022; Kim et al., 2014; Kishore et al., 2019; Smeding et al., 2009). However, previous studies aimed at identifying any possible pre-surgery predictors of post-surgery cognitive decline accepting high false positive error rates **and effect sizes inflation** in the process. In this study, we complement prior findings by identifying a sparse solution to the problem of identifying pre-surgery cognitive profile that is predictive of long-term post-surgery cognitive decline in naturalistic clinical settings. In other words, we aim to describe a minimal significant pre-surgery cognitive profile that predicts higher rate of post-surgery cognitive decline

in a sample derived from everyday clinical practice.

In a typical observational study aiming to determine pre-surgery risk factors of post-surgery cognitive decline, authors employ the following two-step procedure. In the first step, a series of separate univariate analyses for each potential predictor is conducted to pre-select variables for further analysis. In the second step, predictors that achieved an arbitrary threshold (e.g., $p < 0.05$) are used to predict the cognitive decline in a subsequent multiple regression model (Bove et al., 2020; Gruber et al., 2019; Kim et al., 2014; Smeding et al., 2009). This procedure **was named "univariable screening" in biomedical statistics literature and was shown to systematically overestimate effect sizes (Zwet, 2019; Zwet & Cator, 2021). In this article, we show that it** can **also** lead to false positive error rates that are magnitudes higher than the expected nominal five percent.

One way to overcome shortcoming**s of univariable screening is to use partially pooling estimators of the effects associated with each predictor. This can be achieved via regularizition of multiple regression coefficients (Zwet, 2019). In this article, regularizing technique of choice is** the Bayesian Lasso regression, a method developed for identifying small amount of significant predictors out of a larger pool of possible predictors such as results from a comprehensive neuropsychological battery (Park & Casella, 2008).

Another way to achieve sparsity in prediction of post-surgery cognitive decline is to reduce the number of potential predictors. In the context of neuropsychological assessment this can be accomplished straightforwardly via a latent variable approach such as factor analysis that statistically extracts commonalities across several cognitive tasks. Added benefit of employing such a procedure to pre-surgery predictors is that latent variable approaches can reduce the impact of the task impurity problem – the observation that any cognitive task involves several cognitive functions at once (Burgess, 2014; Whitney & Hinson, 2010).

Overall, in this study we aimed to **describe annual post-surgery cognitive decline on group- and patient-level as well as** derive a sparse solution to the task

of identifying pre-surgery cognitive profile predictive of long-term post-surgery cognitive decline in STN DBS treated PD patients. In other words, instead of identifying any pre-surgery cognitive variables that can be predictive of post-surgery decline, we aimed to identify only the most likely predictive ones. To this end, we asked the following research questions: *RQ1)* What is the size of expected long-term rate of cognitive decline after STN DBS in PD patients? *RQ2)* What is the pre-surgery cognitive profile that is predictive of long-term post-surgery cognitive decline in STN DBS treated PD **patients**? To answer these questions, we analyzed data of retrospectively sampled longitudinally followed STN DBS treated PD patients with a single pre-surgery comprehensive neuropsychological assessment and up to five post-surgery cognitive screening assessments.

Materials and methods

Participants

The data of all patients diagnosed with idiopathic PD following United Kingdom Parkinson's Disease Society Brain Bank Criteria (Hughes et al., 1992) that underwent cognitive evaluation for STN DBS treatment at General University Hospital in Prague between years 2000 and 2020 were retrospectively gathered from clinical records and considered for inclusion in the study. **Patients were selected for DBS treatment via criteria mirroring the CAPSIT protocol (Defer et al., 1999), consequently p**atients with atypical parkinsonian syndromes, dementia, depression at the time of pre-surgery assessment (according to an independent psychiatric evaluation), recurrent psychotic conditions or a gait disorder despite optimal dopaminergic therapy during pre-surgery assessment were not implanted and were thus not included in the study. Furthermore, only patients who underwent pre-surgery and at least one post-surgery assessment were included. All included patients were treated via continuous bilateral STN DBS in conjunction with dopaminergic therapy. Bilateral STN DBS implantation was performed as previously described (Jech et al., 2006; Jech et al., 2012; Urgosik et al., 2011). All patients provided signed informed consent and the study was approved by the General University Hospital Ethics Committee in

Prague, Czech Republic.

Assessments

All patients underwent a comprehensive pre-surgery assessment including neuropsychological and neurological examinations. The patients were followed up post-surgery with similar examination protocol at varying time intervals according to their options. Post-surgery, patients were first contacted one year after the surgery and every two years afterwards. The pre-surgery assessment was performed with the usual dopaminergic therapy (ON medication). In the post-surgery assessment, patients were examined in the ON medication condition and STN DBS ON with optimal stimulation parameters.

Pre-surgery neuropsychological measures

The neuropsychological assessment was arranged analogously to the standard International Parkinson and Movement Disorder Society (MDS) neuropsychological battery at Level II for mild cognitive impairment in Parkinson's disease (PD-MCI) (Bezdicek, Sulc, et al., 2017; Litvan et al., 2012). The battery consisted of 10 tests in 5 cognitive domains: (i) attention: Trail Making Test, part A (TMT-A) (Bezdicek et al., 2012; Bezdicek, Stepankova, et al., 2017; Partington & Leiter, 1949) and dot color naming condition from Prague Stroop Test (PST-D) (Bezdicek, Lukavsky, et al., 2015) for sustained visual attention; (ii) executive functions: Trail Making Test, part B (TMT-B) (Bezdicek et al., 2012; Bezdicek, Stepankova, et al., 2017; Partington & Leiter, 1949) for set shifting, and Tower of London task (TOL) (Michalec et al., 2017; Shallice, 1982) for planning; (iii) language: Similarities (Sim.) from Wechsler Adult Intelligence Scale, third revision (WAIS-III) (Wechsler, 2010) for conceptualization, and category verbal fluency test (CFT, category Animals) (Nikolai et al., 2015) for speeded word production; (iv) working memory: Digit Span backward (DS-B) from WAIS-III (Wechsler, 2010) and Spatial Span backward (SS-B) from Wechsler Memory Scale, third edition (WMS-III) (Wechsler, 2011) for auditory and spatial working memory respectively; and (v) memory: Rey Auditory Verbal Learning Test delayed recall (RAVLT-DR) (Bezdicek et al., 2014; Frydrychová et al., 2018) for explicit verbal

learning and memory, and WMS-III Family Pictures delayed recall (FP-DR) for visuo-spatial memory (Wechsler, 2011). Furthermore, we administered the following tests beyond the battery: Prague Stroop Test, naming color of neutral words (PST-W) and interference condition (i.e., naming color of contrasting color words, PST-C) for sensitivity to interference (Bezdicek, Lukavsky, et al., 2015), Controlled Oral Word Association Test (COWAT, letters K + P) (Nikolai et al., 2015) for mental flexibility, and WMS-III letter-number sequencing (LNS) (Wechsler, 2011) for working memory. Finally, anxiety was assessed with the State-Trait Anxiety Inventory for the state (STAI-X1) and trait (STAI-X2) anxiety (Spielberger et al., 1983).

Longitudinal neuropsychological measures

Patients' longitudinal cognitive state was assessed pre-surgery and post-surgery with MDS battery at Level I using Mattis Dementia Rating Scale, second edition (DRS-2) (Bezdicek, Michalec, et al., 2015; Jurica et al., 2001). DRS-2 is a routinely employed cognitive screening measure in PD that has been shown to have acceptable discriminative performance for PD-MCI in Czech population with both sensitivity estimated to be around 0.8 (Bezdicek, Michalec, et al., 2015; Mazancova et al., 2020). Furthermore, subjective depressive symptoms were assessed with Beck Depression Inventory, second edition (BDI-II) (Beck et al., 1996; Ciharova et al., 2020) at each assessment. BDI-II was not used for pre-surgery exclusion due to depression which was instead ascertained by an independent neuropsychiatric evaluation.

Neurological examination

Patients' motor state was assessed with part three of the Movement Disorders Society Unified Parkinson's Disease Rating Scale (MDS-UPDRS III) in medication ON and medication OFF state during the pre-surgery levodopa test. Scores of patients who underwent the older version of the Unified Parkinson's Disease Rating Scale (UPDRS III) were converted to the MDS-UPDRS III scale using the method described by Hentz et al. (2015). The levodopa equivalent daily dose (LEDD) was calculated at each assessment time-point according to Tomlinson et al. (2010).

Theoretical and empirical estimands

Following the framework of Lundberg et al. (2021), in this section we link our research questions to explicit targets of inference (i.e., theoretical estimands) and to observable data (i.e., empirical estimads). Within this framework, the theoretical estimand consists of two components, a unit-specific quantity and the target population. Regarding our *RQ1*, the patient-specific quantity of interest is the difference between expected post-surgery cognitive performance and expected cognitive performance *k* years before where *k* can be any positive real number. We aimed to describe this quantity on two levels: (i) the current sample and (ii) a population of patients selected for DBS treatment via the CAPSIT-protocol criteria. To allow for generalization beyond our sample as per the second version of this theoretical estimand, we assume exchangeability between patients selected via CAPSIT criteria to the extend that can be quantified by patient-level variance estimated from our sample (see Yarkoni, 2020). The patient-specific quantity for *RQ2* is the difference between expected post-surgery cognitive decline of a patient with fixed level of pre-surgery performance across all cognitive variables and expected post-surgery cognitive decline of patients with performance that is one unit smaller in a single cognitive variable but equal to this patient's performance otherwise. This quantity was described for the sample only. Empirical estimands were the same quantities, conditional on patient being selected for the study (based on geographical and exclusion criteria described above). Importantly, all three estimands are descriptive, not causal.

Statistical analyses

Deriving pre-surgery cognitive profile

Latent cognitive factors were extracted from the data via an exploratory factor analysis (EFA) with varimax rotation using ordinary least squares to find the minimum residual solution (Harman & Jones, 1966). We opted for the orthogonal varimax

rotation because: (i) extracting orthogonal factors can be statistically advantageous in later steps of our analysis due to reducing multicollinearity, and (ii) in the framework of PD-MCI, it is considered desirable to describe patients' cognitive profile by factors or tests that are independent of each other (Litvan et al., 2012).

All pre-surgery cognitive tests listed above were entered into EFA as input variables (see **the** Supplementary material for the exact processing pipeline). Missing observations were multiply imputed using a parametric bootstrap via the "missMDA" R package to create one hundred imputed data sets. We then computed EFA with three up to eight factors via the "psych" R package (Josse & Husson, 2016; R Core Team, 2024; Revelle, 2022) using each imputed data set. Within each imputed data set, factor scores for each patient were calculated using the regression method (Thomson, 1951).

We based the number of extracted factors on a combination of the root-mean-square error approximation (RMSEA), Tucker-Lewis Index (TLI), and consistency of each factor model across imputations. TLI is a measure of a goodness-of-fit such that higher values of TLI imply better fit and values exceeding 0.90 are considered to indicate a good model fit. On the other hand, RMSEA is a measure of badness-of-fit such that lower values imply better fit with values less than 0.08 indicating an adequate model fit (Browne & Cudeck, 1992). A model was considered consistent if **it** identified similar factors across imputed data sets.

Describing and predicting post-surgery cognitive decline

Longitudinal data were analyzed using Bayesian generalized linear mixed models (GLMMs). GLMMs overcome th**e** issue **of confounding measurement error with true change plaguing change scores analysis** by estimating both group-level (i.e., "fixed effect") as well as patient-level (i.e., "random effect") parameters. Furthermore, modelling patient-level effects results in partial pooling of parameter estimates (shifting parameter estimates towards each other), which reduces the influence of outliers and facilitates reliable group-level inference (Gelman et al., 2012; Tuerlinckx et al., 2006).

To describe the rate of post-surgery cognitive decline, we estimated a GLMM with longitudinal DRS-2 performance as an outcome predicted by the time after surgery

on the group-level and correlated patient-specific intercepts and slopes on the patient-level. Since the group-level slope of this model represents the expected rate of cognitive decline after STN DBS, it constituted **statistical estimate of the sample version of our** *RQ1* estimand **(i.e., the expected annual cognitive decline in the sample). To arrive at statistical estimate of the population version of our *RQ1* estimand (i.e., the expected annual cognitive decline in a population of patients selected for surgery using CAPSIT-protocol criteria) we used the model to predict expected post-surgery cognitive decline at one year post-surgery intervals compared to a pre-surgery assessment using both group- and patient-level parameters. We** evaluate**d** suitability of the linear model **by** compar**ing** it to an **otherwise** equivalent non-linear model that estimated post-surgery cognitive trajectory via tensor product smooths (Wood et al., 2012). Both models were fitted using non-informative improper flat priors **for group-level parameters** to ensure that their parameters are informed primarily by the data.

T**o evaluate predictive utility of pre-surgery cognitive profile,** we **estimated further two** GLMMs. **L**ongitudinal DRS-2 performance was predicted on a group-level by post-surgery time slopes varying by either patients' pre-surgery cognitive tests' scores (the "test scores" model) or patients' pre-surgery latent cognitive factors' scores extracted from the EFA reported above (the "factor scores" model). Both models further included correlated patient-level intercepts and slopes. To check robustness of our findings we compared the results to estimates of GLMMs that also included group-level effects of age, LEDD and BDI-II (and their interaction with time after surgery).

Since previous long-term studies demonstrated that a subset of PD patients treated with STN DBS can develop dementia which may lead to heavy tails in the data distribution of cognitive test scores, we **used a** Student-t instead of Gaussian **measurement error model**. Furthermore, because the outcome DRS-2 has a maximum of 144 points which is achieved by a large proportion of healthy people (Bezdicek, Michalec, et al., 2015), we used the right-censored version of Student-t to

account for the ceiling effect. Models' likelihoods had following specification:

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

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

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

$i = 1 \ldots n$, where n is the total number of assessments across all patients, m is the total number of pre-surgery predictors, DRS_{max} is the maximal attainable score in DRS-2 (i.e., a raw score of 144), *T()* is the Student-t cumulative distribution function, *t()* is the Student-t probability density function, $time_i$ is the time from surgery at assessment i, $predictor_{[j]i}$ is the pre-surgery cognitive score in the predictor (i.e., either a test or latent factor) j of the patient evaluated at assessment i, τ **parameters represent patient-level variance,** z_k **and** x_k **represent standardized patient-level effects for** k^{th} **patient,** μ_i **represents true score estimate of** k^{th} **patient's cognitive performance on assessment at** $time_i$, and remaining terms denote model parameters. Estimands relating to *RQ2* comprised of the two sets of $\delta_{predictor[j]}$ representing the expected prognostic value of single pre-surgery cognitive tests and latent cognitive factors.

We specified equivalent prior distributions for model parameters of both the "test scores" and the "factor scores" models. We used the Bayesian Lasso priors for all group-level parameters barring the intercept. This prior is the Bayesian equivalent of the Lasso method for performing variable selection and allows for fitting models with a large number of potentially collinear predictors. All remaining parameters were given weakly informative priors to ensure that models' estimates fall within the range of measurable values of the outcome.

Exploring association of stimulated STN site with cognitive decline

Although our primary goals relate to the purely descriptive task of predicting post-surgery cognitive decline, further studies might want to ask causal questions about possible interventions to change the rate of this

decline. A straightforward way to possibly affect cognitive performance after STN DBS surgery is to change active stimulation contacts such that volume of affected tissue (VAT) overlaps with motor component of STN while avoiding associative and limbic STN components because high frequency stimulation that improves motor symptoms can result in adverse cognitive side effects (David et al., 2020). Even though our retrospective sample did not contain enough high-quality data that would be needed to reliably evaluate association between STN sites affected by DBS and post-surgery cognitive decline, we provide an exploratory analysis that can provide pointers for future investigations.

For this analysis, we used a subsample of patients for whom we were able to retrospectively obtain (i) pre-surgery magnetic resonance image (MRI) for STN localization, (ii) post-surgery MRI for electrode localization, and (iii) stimulation parameters at the time of MRI assessment for VAT computation. Lead-DBS software (Horn et al., 2019; Horn & Kühn, 2015) was utilized to determine the position of DBS leads and active contacts with DISTAL subcortical atlas for STN compartmentalization (Ewert et al., 2018). The overlap of VAT and the entire STN as well as its motor, associative, and limbic components separately was calculated, providing four overlap volumes for each side. For each patient with VAT available, only one set of overlaps was estimated which corresponded to the time of MRI acquisition, not the time of cognitive assessment as these were not always aligned. We then explored association between VAT overlaps with STN components and post-surgery cognitive decline using a model analogous to the predictive GLMMs described above. For more detailed information about MRI acquisition parameters and modelling choices regarding this analysis see the Supplementary material.

Model description and statistical testing

Estimates were described by **full posterior distributions,** medians and 95% highest density posterior probability intervals (PPIs) of corresponding model parameters **or predictions as appropriate**. A 95% PPI can be interpreted **as the narrowest interval** such that a given parameter **or prediction lies within it with 95% probability. In one case, when presenting results for the second version of our *RQ1* estimand, we report medians and 90% equal-tailed posterior probability intervals (ETIs) instead. A 90% ETI can be interpreted such that a given parameter or prediction lies with 5% probability above its upper bound and with 5% probability below its lower bound.**

Models were compared via the expected log pointwise predictive density (ELPD) computed via the leave-one-out cross-validation (LOO-CV) as approximated by the Pareto-smoothed importance sampling (PSIS) (Vehtari et al., 2015). The ELPD difference ($ELPD_{dif}$) and its 95% frequentist confidence interval (CI) were used to decide whether predictive performance of compared models statistically significantly differs (i.e., the 95% CI excludes zero). To identify influential observations, we calculated a Pareto-k diagnostic and looked for observations with Pareto-k > 0.7 which can be considered problematic (Bürkner et al., 2020; Vehtari et al., 2015).

Evaluating false positive error rates

To validate the assumption that our analysis provides lower false positive rates than the commonly used **univariable screening** procedure, we conducted **a** series of simulations with data set structure equivalent to that observed in our data. Patients' outcome was generated as a normally distributed random variable with unit standard deviation and mean depending on average annual rate of cognitive decline and patient-specific random deviations. Moreover, for each patient we generated a set of potential predictors including either seven independent variables, twenty-three independent variables or twenty-three covaried variables representing our analysis of the predictive utility of seven latent cognitive factors and twenty-three observed cognitive test scores respectively. Covariance structure in the case of covaried predictors was

based on the structure of the battery described above with predictors that represented test measures belonging to the same superordinate task having Pearson's correlation of 0.7 (thus sharing approximately half of the variance) and zero otherwise (see Figure S**11** in **the** Supplementary material). The simulations were set-up such that there was no effect of any predictor on the outcome. Subsequently, we generated one hundred data sets which were then fitted via the **univariable screening** procedure and **the** Bayesian Lasso. For each procedure, number of statistically significant interactions between time and any of the predictors were recorded to estimate the amount of false positive errors.

Transparency and openness

All GLMMs were fitted using via Stan's (version 2.21.0) build-in Hamiltonian Monte Carlo sampler accessed via R version **4.3.3** using package "brms" (Bürkner, 2017; R Core Team, 2024; Stan Development Team, 2020). Four parallel chains were run each for 2,500 iterations for each GLMM. The first 500 iterations served as a warm-up and were discarded. Convergence was checked numerically by inspection of the $\hat{R}s$ and visually by inspection of trace plots. We used R packages "tidyverse" and "dplyr" for data operations, "tidybayes" for operations with model posteriors, and "DiagrammeR," "ggplot2," **"ggridges,"** and "patchwork" for plotting (Iannone, 2022; Pedersen, 2020; Wickham, 2016; Wilke, 2024). This study's design and its analysis were not pre-registered. The data are not publicly available due to privacy or ethical restrictions. The computer code used in our data analysis as well as synthetic data and rep**roducible** code for simulations to estimate false positive error rates can be accessed at https://github.com/josefmana/dbs_cogPRED.

Results

Characterizing the sample

A total of 200 patients with PD who underwent cognitive evaluation for STN DBS between 2000 and 2020 were identified by a retrospective search of local database in General University Hospital in Prague and a total of 126 patients met inclusion criteria (see Figure 1). All included patients were Caucasians and were speaking Czech as their primary language. Baseline demographic and clinical characteristics as well as

stimulation parameters of the sample are presented in Table 1 and baseline cognitive characteristics are presented in Table 2. Mean duration of a follow-up after the surgery was 3.54 years (SD = 2.32, median = 3.07, range = 0.72–11.38) with a median number of 3 assessments per patient (range = 2–6) (see also Figure 2).

Table 1

Clinical characteristics of the sample of included patients

	N	Md	Min-Max	M	SD
Baseline characteristics					
Age at surgery (years)	126	58	40-76	57.25	7.96
Education (years)	117	13	10-23	14.26	2.91
Sex (males)	83 (66 %)	-	-	-	-
Disease duration at surgery (years)	125	11	4-30	11.67	4.05
LEDD (mg)	114	1614	400-4138	1696.88	672.33
Levodopa test (% response)	93	54	20-81	52.64	12.81
MDS-UPDRS III (ON medication)	105	21	7-46	21.78	7.57
MDS-UPDRS III (OFF medication)	100	45	24-81	45.79	10.93
Stimulation parameters[1]					
Current right (mA)	67	2.1	0.6-4.3	2.14	0.71
Current left (mA)	67	2.3	1.0-3.9	2.35	0.68
Voltage right (V)	59	3.0	1.4-5.3	3.00	0.65
Voltage left (V)	59	2.9	0.5-5.7	2.87	0.74
Pulse duration right (ţs)	126	60.0	52.0-120.0	73.98	17.14
Pulse duration left (ţs)	126	60.0	30.0-120.0	71.57	16.15
Frequency right (Hz)	126	130.0	60.0-210.0	128.42	12.44
Frequency left (Hz)	126	130.0	60.0-160.0	127.89	11.14

[1]Each measurement of each electrode considered independently. For stimulation parameters, column N indicate number of patients with current/voltage mode of stimulation. N: number of observations; Md: median; M: mean; SD: standard deviation; 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 as described in the main text; V: Volts; mA: milliampere; s: microseconds; Hz: Hertz.

Table 2

Pre-surgery neuropsychological measures of included patients

Test	N	Md	Min-Max	M	SD
DRS-2 (range 0-144)	126	141	129-144	139.77	3.68
BDI-II (range 0-63)	122	8	0-28	9.28	5.95
STAI-X1 (range 20-80)	104	37	23-63	38.27	8.66
STAI-X2 (range 20-80)	104	39	22-62	39.52	8.11
TMT-A (secs)	125	41	18-122	43.15	15.85
TMT-B (secs)	124	102	39-334	119.01	54.96
DS-F (range 0-16)	113	8	5-16	8.94	2.02
DS-B (range 0-14)	113	6	2-11	6.21	1.80
LNS (range 0-21)	97	8	2-13	7.85	2.46
SS-F (range 0-16)	110	8	4-14	7.54	1.74
SS-B (range 0-16)	110	7	2-11	6.97	1.69
TOL (range 0-108)	118	78	46-90	74.93	9.81
PST-D (secs)	124	13	8-20	13.09	2.37
PST-W (secs)	124	15	10-25	15.72	2.97
PST-C (secs)	124	28	14-57	29.35	9.15
COWAT (total words)	125	32	12-57	32.35	9.05
CFT (words/min.)	89	22	3-39	22.55	7.10
Sim. (range 0-28)	94	22	8-28	21.61	4.35

RAVLT-IR (range 0-75)	108	44	20-64	43.80	8.39
RAVLT-B (range 0-15)	108	5	0-8	4.71	1.45
RAVLT-DR (range 0-15)	108	8	3-14	8.37	2.49
RAVLT-Rec50 (range 0-50)	105	46	33-50	45.10	3.49
RAVLT-Rec15 (range (0-15)	107	14	9-15	13.32	1.54
FP-IR (range 0-64)	74	32	15-55	32.04	10.21
FP-DR (range 0-64)	74	32	13-55	31.91	9.97

N: number of observations; Md: median; M: mean; SD: standard deviation; DRS-2: Dementia Rating Scale, second edition; BDI-II: Beck Depression Rating Scale, second edition; 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; Secs: seconds; Total words: word count in two minutes (one minute per each letter P and K); words/min.: word count in one minute time limit.

Pre-surgery cognitive profile

Detailed summaries of the fit statistics of all EFA models are presented in the Supplementary material (see Table S1 and Figure S1). Most importantly, raising the number of factors from six to seven resulted in a clear improvement. Out of the one

hundred imputed data sets, the six-factor model showed good fit according to RMSEA in 96 cases and it showed good fit according to the TLI in 76 cases. On the other hand, the seven-factor model showed good fit according to RMSEA in 99 cases and good fit according to TLI in 97 cases. Moreover, the seven-factor model was more consistent across imputations. Finally, while the eight-factor resulted in the best fit statistics, factors identified by this model were often substantially loaded on by only a single cognitive test score (with a factor loading above 0.3) which impedes theoretical interpretation of such factors. Consequently, the seven-factor model was retained for subsequent analyses. On average, the seven factors accounted for a total of 54.8 % of variance (SD = 0.8 %) and corresponded to seven cognitive functions: 1) executive functions/attention (EF/Att.) was loaded on primarily by PST tasks, TMT tasks, verbal fluency tests and TOL, 2) episodic memory (EM) was loaded on primarily by indexes of RAVLT except for the recall of interference list (RAVLT-B), 3) verbal working memory (VWM) was loaded on primarily by Digit Span tasks, LNS and Similarities, 4) visuospatial memory (VM) was loaded on primarily by indexes of the Family Pictures test, 5) set shifting (SS) was loaded on primarily by TMT tasks and RAVLT-B, 6) anxiety (An.) was loaded on primarily by STAI, and 7) spatial working memory (SWM) was loaded on primarily by Spatial Span tasks (see Table 3).

Table 3

Summary of factor loadings

	EF/Att.	EM	VWM	VM	SS	An.
TMT-A	-0.26 (0.10)	0.05 (0.03)	0.09 (0.08)	-0.16 (0.11)	-0.38 (0.28)	-0.08 (0.05)
TMT-B	-0.30 (0.09)	-0.08 (0.04)	-0.31 (0.10)	-0.12 (0.11)	-0.53 (0.20)	-0.02 (0.05)
DS-F	0.05 (0.04)	-0.01 (0.03)	0.67 (0.09)	-0.04 (0.05)	-0.02 (0.09)	-0.08 (0.03)
DS-B	0.27 (0.05)	0.12 (0.03)	0.63 (0.07)	-0.05 (0.07)	0.05 (0.08)	0.08 (0.03)
LNS	0.19 (0.06)	0.06 (0.04)	0.49 (0.10)	0.14 (0.10)	0.20 (0.10)	-0.21 (0.09)
SS-F	0.05 (0.04)	0.06 (0.03)	0.19 (0.09)	0.10 (0.14)	0.03 (0.14)	-0.12 (0.06)
SS-B	0.04 (0.04)	-0.06 (0.04)	0.18 (0.09)	0.32 (0.12)	0.13 (0.09)	-0.26 (0.07)

TOL	0.38 (0.07)	0.02 (0.04)	0.07 (0.06)	0.14 (0.08)	0.25 (0.22)	0.06 (0.05)
PST-D	-0.76 (0.04)	-0.01 (0.03)	-0.11 (0.05)	-0.06 (0.05)	-0.19 (0.09)	-0.02 (0.03)
PST-W	-0.83 (0.04)	-0.05 (0.03)	-0.30 (0.05)	-0.14 (0.05)	-0.06 (0.09)	0.01 (0.04)
PST-C	-0.52 (0.06)	-0.06 (0.03)	-0.27 (0.05)	-0.14 (0.07)	-0.18 (0.15)	-0.01 (0.04)
COWAT	0.43 (0.06)	0.24 (0.04)	0.12 (0.04)	0.10 (0.09)	0.14 (0.09)	0.04 (0.05)
CFT	0.37 (0.10)	0.24 (0.05)	-0.05 (0.09)	0.30 (0.09)	0.28 (0.13)	-0.12 (0.06)
Sim.	0.14 (0.05)	0.10 (0.04)	0.48 (0.08)	0.04 (0.08)	0.16 (0.12)	-0.09 (0.07)
RAVLT-IR	0.26 (0.05)	0.79 (0.04)	0.09 (0.06)	0.03 (0.07)	0.13 (0.07)	0.05 (0.04)
RAVLT-B	0.15 (0.05)	0.18 (0.04)	0.28 (0.08)	0.13 (0.08)	0.38 (0.26)	-0.03 (0.05)
RAVLT-DR	0.05 (0.04)	0.74 (0.03)	-0.02 (0.04)	0.07 (0.05)	0.04 (0.06)	0.12 (0.03)
RAVLT-Rec50	0.07 (0.04)	0.68 (0.03)	0.15 (0.05)	0.11 (0.06)	-0.04 (0.07)	-0.02 (0.03)
RAVLT-Rec15	-0.11 (0.04)	0.48 (0.04)	0.04 (0.05)	0.24 (0.09)	0.02 (0.06)	-0.04 (0.05)
FP-IR	0.22 (0.09)	0.30 (0.08)	-0.01 (0.09)	0.70 (0.14)	0.13 (0.08)	-0.13 (0.08)
FP-DR	0.24 (0.07)	0.26 (0.08)	-0.01 (0.08)	0.72 (0.13)	0.14 (0.07)	-0.13 (0.08)
STAI-X1	0.00 (0.03)	-0.05 (0.03)	-0.09 (0.05)	-0.05 (0.06)	-0.03 (0.04)	0.79 (0.13)
STAI-X2	0.05 (0.03)	0.14 (0.04)	-0.04 (0.04)	-0.12 (0.07)	0.06 (0.05)	0.69 (0.10)
Proportion Var	0.11 (0.02)	0.10 (0.01)	0.08 (0.01)	0.07 (0.01)	0.06 (0.02)	0.06 (0.01)
Cumulative Var	0.11 (0.02)	0.21 (0.02)	0.30 (0.02)	0.37 (0.02)	0.43 (0.02)	0.49 (0.02)

Values represent mean (SD) across one hundred imputations. Factor loadings used for interpretation (|loading| $>$ 0.30) are printed in bold. 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; STAI-X1: State-Trait Anxiety Inventory, the state version; STAI- X2: State-Trait Anxiety Inventory, the trait version; Secs: seconds; Total words: word count in two minutes (one minute per each letter P and K); words/min.: word count in one minute time limit. Proportion Var: Proportion of variance in data accounted for by each factor (column); Cumulative Var: Cumulative variance accounted for by each factor and factors that preceded it (columns to the left); EF/Att.: Executive functions/Attention; EM: Episodic memory; VWM: Verbal working memory; VM: Visuospatial memory; SS: Set shifting; An: Anxiety; SWM: Spatial working memory.

Describing post-surgery cognitive decline

Both descriptive longitudinal GLMMs converged within a specified number of iterations ($\hat{R}s \leq 1.01$). All observations had Pareto-k below 0.7 implying that the results are not likely biased by influential outliers. The linear and non-linear models showed tight correspondence up to approximately five years post-surgery after which the non-linear model predicted a slightly faster rate of cognitive decline than the linear model (see Figure 3). The difference in estimated predictive performance between these

models did not reach statistical significance ($ELPD_{dif}$ = 1.64, 95% CI [-2.13, 5.41]). Based on the linear model, there was an average post-surgery decline of 0.90 DRS-2 points/year (95% PPI [-1.19, -0.62]) from an average pre-surgery DRS-2 performance of 140.34 out of 144 points (95% PPI [139.61, 141.07]).

In Table 4, we present expected true score differences as well as model's predictions for new observations at five yearly post-surgery assessment times compared to pre-surgery assessment derived from group-level parameters only (reflecting the sample version of our *RQ1* estimand), both group- and patient-level parameters (reflecting the population version of our *RQ1* estimand), and the full model with both true score and measurement error. Although median expected post-surgery cognitive decline was similar across versions, the ETIs of the population estimates and full model's predictions indicate that a large proportion of observed variability in post-surgery cognitive decline is due to inter-individual heterogeneity and measurement error respectively (see also Figures S3 and S4 in the Supplementary material).

Table 4

Posterior predictions of cognitive changes after surgery

	True score predictions		
	Group-level parameters[a]	Group- & Patient-level parameters[b]	Observed scores pred
Yearly decline[d]			
Intercept	140.34 [139.71, 140.95]	140.35 [135.71, 144.00]	140.37 [132.97, 14
Slope	-0.90 [-1.14, -0.67]	-0.78 [-2.41, 0.52]	-0.73 [-9.24, 7.6
Contrasts			
Y1-minus-Pre	-1.17 [-1.49, -0.87]	-1.03 [-3.14, 0.67]	-1.09 [-9.45, 7.4
Y2-minus-Pre	-2.08 [-2.63, -1.55]	-1.87 [-5.60, 1.18]	-2.00 [-10.11, 7.
Y3-minus-Pre	-2.98 [-3.77, -2.22]	-2.71 [-8.11, 1.67]	-2.84 [-11.12, 6.

Y4-minus-Pre	-3.88 [-4.92, -2.89]	-3.56 [-10.65, 2.13]	-3.76 [-12.04, 5.
Y5-minus-Pre	-4.79 [-6.06, -3.56]	-4.41 [-13.23, 2.55]	-4.71 [-12.92, 5.

[a]Contrasts for the sample version *RQ1* estimand predicted by $\mu_i \sim \alpha + \delta_{time} time_i$

[b]Contrasts for the population version *RQ1* estimand predicted by $\mu_i \sim \alpha + \delta_{time} time_i + \alpha_{id[i]} + \delta_{id[i]} time_i$

[c]Contrasts for model's prediction of the raw score sampled from $t(\vartheta, \mu_i, \sigma)$

[d]The rows represents expactation of patients' performance at pre-surgery assessment, i.e., 0.3 years before surgery (Intercept), and expected annual Dementia Rating Scale decline (Slope).

Yi: assessment i years post-surgery; values represent posterior prediction median [90% equal tailed interval (ETI)]; we used 90% ETI instead of the 95% highest density posterior predictive intervals (PPIs) used elsewhere in the article because 90% ETI can be interpreted such that there is 5% probability of observing value smaller than its lower bound and 5% probability of observing value bigger than its upper bound which may not hold for PPIs; all values were calculated by first generating predcitions from the linear desciptive models using parameters specified above and then censoring values above 144 or below 0 before calculating medians and 90% ETIs.

Predicting post-surgery cognitive decline

Both predictive longitudinal GLMMs converged within a specified number of iterations ($\hat{R} \leq 1.02$) with all observations having Pareto-k below 0.7. **Including either pre-surgery cognitive tests or pre-surgery cognitive factors as predictors notably decreased patient-level inter-individual heterogeneity compared to the time only model even though all three models generated similar predictions for majority of included patients (see Figures S2 and S4 in the Supplementary material)**

Patients with lower verbal working memory or set shifting showed relatively impaired pre-surgery **DRS-2** performance while there was no cognitive test that clearly indicated pre-surgery **DRS-2** impairment performance (see Tables S2 and S3 in **the** Supplementary material). Patients with lower pre-surgery executive functions/attention

performance showed faster post-surgery cognitive decline (see Figure 4). Pre-surgery executive functions/attention performance that was one standard deviation below sample average was associated with additional 0.39 DRS-2 points post-surgery annual decline (95% PPI [-0.63, -0.15]). There was no single cognitive test that clearly indicated faster-than-average post-surgery cognitive decline (all 95% PPIs included zero**, see Table S2**). Adding group-level effects of age, LEDD and BDI-II did not reveal any substantial deviation from these results (see Figures S**5** and S**6** in **the** Supplementary material).

Electrode localization exploration

Total of 69 patients was included into analysis (see the Supplementary material Figure S7 for electrode localization visualization, and Tables S4 and S5 for description of this subsample). Patients included in this analysis were marginally younger with less pre-surgery anxiety, less depressive symptoms and better performance in some measures of attention and executive functions (Tower of London task, Category Fluency Test and Trail Making Test part A) compared to patients excluded from this analysis (see Figure S8 in the Supplementary material).

Overall, there was no statistically clear evidence that proportions of STN components affected by DBS are associated with the degree of post-surgery cognitive decline in our sample (see Figure S9 and Table S6 in the Supplementary material). However, some patterns can be observed in our results. Although associations of post-surgery cognitive decline with proportion of affected motor and limbic STN components were all centred around zero, there was clearly more uncertainty in estimates regarding limbic STN compared to motor STN. Moreover, there was specifically higher (negative) posterior collinearity of limbic STN components time-dependent parameters with other time-dependent parameters (see Figure S10 in the Supplementary material). Finally, our results show weak trend towards potential association between proportion of affected right

associative STN and post-surgery cognitive decline. According to our model and data, there was almost 80% probability that patients with higher VAT/right associative STN overlap show faster post-surgery cognitive decline than patients with lower overlap.

Evaluating false positive error rates

Results of simulations used to estimate false positive rates of the **univariable screening** procedure and the Bayesian Lasso are summarised in **the** Supplementary material Figure S**12**. Overall, the Bayesian Lasso showed almost no false positives across simulation settings whereas the false positive rates of the **univariable screening** procedure ranged from 14 to 57 of analyses including at least one false positive. In the case of our data structure, the false positive rates were attenuated when all twenty-three predictors covaried or when we reduced the number of predictors to seven independent variables.

Discussion

In the present study, we analyzed retrospectively sampled data of longitudinally followed 126 PD patients after STN DBS surgery and described their post-surgery cognitive performance. We observed a mild **average** post-surgery cognitive decline **with considerable inter-individual variability** (*RQ1*) **. The post-surgery decline** was faster in patients with lower pre-surgery executive functions/attention (*RQ2*) compared to the rest of the sample. **Our exploratory analysis of association between affected proportion of STN components and post-surgery cognitive decline did not yield any clear results, however some patterns of our data can be used to guide further research.**

Description of post-surgery cognitive decline

Expected average annual rate of cognitive decline after STN DBS in **our sample of** PD **patients** reache**d** 0.90 from a total of 144 points **in DRS-2. Whereas the sample estimate was tightly clustered around this circa one point a year post-surgery decline, generalization to a population of patients selected via the CAPSIT protocol criteria resulted in estimates of annual post-surgery**

change ranging from 2.5 points decline to 0.5 points improvement with 90% certainty (Table 4). As a reference point, compare these estimates to results of Pedraza et al. (2007) who reported values of 6 and 7 DRS-2 points decline as reliable change cutoffs for European Americans based on 90% predictive intervals for approximately one- and two-years re-tests respectively. As both our sample and population true score change estimates fall above these cut-offs with high level of certainty, we may conclude that STN DBS seems to be relatively safe from cognitive point of view at least up to two years post-surgery.[1]

The estimate of an average annual cognitive change from our study represents **somewhat** slower rate **of decline** than previous reports. **For example, Gruber et al. (2019) reported 1.6 DRS-2 points/year decline (90% CI [-2.74, -0.46]) based on change scores of 32 patients assessed pre-surgery and at various time points post-surgery (compare to the "Slope" row of Table 4). Additionaly, based on one year post-surgery change scores, Smeding et al. (2009) reported decline of 2.4 (90% CI [-3.61, -1.19]) DRS-2 points in 105 patients, and Reich et al. (2022) reported decline of 2.4 (90% CI [-3.63, -1.18]) DRS-2 points in 32 patients (compare to the "Y1-minus-Pre" row of Table 4). Other studies (Boel et al., 2016; Castrioto et al., 2022; Schupbach, 2005) appear to observe post-surgery decline similar to or larger than our estimate, however, since they report neither cognitive decline parameter estimates, nor raw change scores, we were unable to ascertain the (dis)similarity between their and our observations numerically.**

[1] Based on the descriptive linear model's predictions summarized in Table 4, posterior probability of seeing a patient with DRS-2 score change lower than these reliable change cut-offs was respectively 0%, 0% and 14% for the sample true score, population true score and observed score predictions one year after surgery, and 0%, 1% and 13% for the sample true score, population true score and observed score predictions two years after surgery.

Using our model as an interpretation device

Next we demonstrate how our linear descriptive model can be used to inform decisions in future studies by comparing its predictions directly to data reported in Reich et al. (2022). We have selected this article because it represents a prototypical pre-surgery/post-surgery study design for assessing cognition in PD patients treated by STN-DBS, it uses DRS-2 as its outcome, and the raw data (change scores) analyzed for purposes of the article can be readily identified from Figure 4B therein.

To decide whether data such as those evaluated by Reich et al. (2022) constitute true score changes (as is assumed by change scores analysis) or whether there is appreciable measurement error, we can compare the data distribution with the population true score estimate of our model (Table 4, second column). The model implies that not more than 5% of the sample should fall below 3.1 DRS-2 points decline at year one post-surgery assessment if all we observed were true score changes. Inspected data set contains 7/32 (22%) patients who fall below this threshold. On the other hand, only 3/32 (9%) patients fall below our model's bottom 5% expectation based on prediction of true score with added measurement error (Table 4, third column).[2]

Comparison of inspected data set with our model's predictions thus leaves researchers with two main takeaways: (i) the data could have come from a distribution approximated by the model because its raw score predictions align well with the data, and (ii) high amount of change score variance can be attributed to measurement error because we observed

[2] As the data set we are inspecting contains only 32 patients, small percentage predictions might easily be misaligned between model predictions and observed sample. However, the general observation that the sample contains significant amount of outliers with respect to true score change model predictions but not for raw score model predictions holds in our case even for larger intervals such as 80% ETIs (i.e., inspecting bottom 10% instead of 5%) and the conclusion of this section thus seem valid.

significantly larger proportion of extreme values than predicted by the model with measurement error removed. Such an analysis will provide researchers with useful information they can act upon by increasing sample size to offset lost efficiency due to measurement error or adopt analysis procedure that accounts for measurement error explicitly (Gelman & Vákár, 2021; Van Bork et al., 2023). Alternatively, researchers can hypothesize that the discrepancy between model and observed data is not due to measurement error but rather reflects differences between populations. This line of reasoning could motivate researchers to account for population differences via procedures such as post-stratification to improve their estimates (Deffner et al., 2022). Finally, researchers could surmise that our model is inappropriate for describing their data. If this was the conclusion researchers made, they can easily use our model to generate predictions for their sample using model's full posteriors we share in a plain text format on article's repository (https://github.com/josefmana/dbs_cogPRED) and attempt to falsify it.

Predictive pre-surgery cognitive profile

When predicting post-surgery cognitive decline via pre-surgery cognitive profile, we aimed to identify a sparse solution including only these pre-surgery cognitive variables that are the most likely to be truly predictive. To achieve this goal, we applied the Bayesian Lasso and factor analysis to decrease false positive error rates of our analysis and validated effectiveness of these procedures via a set of simulations.

In **our sample**, lower performance on the latent factor of executive functions/attention (EF/Att.) was reliably predictive of post-surgery cognitive decline. Similarly, previous studies suggested that patients with executive deficit (operationalized as performance on tasks such as Stroop test, TMT, Wisconsin Card Sorting Test and verbal fluency test) are at a high risk of developing dementia and experiencing fast cognitive decline after STN DBS surgery (Bove et al., 2020; Kishore et

al., 2019; Smeding et al., 2009). **On the other hand, recent meta-analysis (Jahanshahi et al., 2022) identified both pre-surgery executive dysfunction as well as poorer pre-surgery memory to be predictive of post-surgery cognitive decline, the latter of which was not replicated in our study.**

Although there is a large body of evidence showing that pre-surgery executive deficit is predictive of post-surgery cognitive decline, it is unclear which executive functions components provide the most information. In our study, the EF/Att. factor was loaded on primarily by timed test scores and could be specific to processing speed component of executive functions. Conversely, the set shifting factor which represent another executive functions' component did not clearly predict post-surgery cognitive decline above and beyond the remaining pre-surgery cognitive variables in our data set. However, the set shifting factor part of our *RQ2* estimand showed less consistency across imputations in its derivation (see high standard deviations in Table 3) and high variability in its predictive value (see Figure 4). We thus cannot give any firm conclusion regarding set shifting's role for post-surgery cognitive decline prediction. Future studies could benefit from further differentiating executive functions to identify the best predictive executive components.

The role of STN sites affected by DBS

Our exploratory analysis of post-surgery cognitive decline prediction by proportions of STN components being affected by DBS yielded equivocal results. As our data sampling design was not optimized for evaluation of direct stimulation outcomes' effect on post-surgery cognitive decline, this estimation inefficiency likely reflects several sources of noise that could be better controlled for in future studies. On top of the outcome measurement error discussed above, these noise sources include alignment between time of cognitive assessment and electrode localization, and reliability of localization itself. In our study, only a single electrode localization was derived for each

patient and the time gap between MRI acquisition and cognitive assessment differed between patients and assessment times. Our results would thus be most accurate under the unrealistic assumption that proportion of STN being affected by DBS is time-invariant across patients. This assumption can be relaxed by gathering repeated MRI data for electrode localization at times corresponding to patients' cognitive evaluations data. Furthermore, although electrode localization via Lead-DBS has been shown to be relatively reliable and comparable across raters and modalities (Lofredi et al., 2022; Xu et al., 2023), using data from several raters and employing post-surgery computational tomography instead of MRI may increase precision of electrode localization estimates.

Constraints on Generality

Several not yet discussed constraints of generality apply to our findings. Due to the lack of a control group, we cannot discern the causal effect of DBS from the effect of disease progression. Consequently, we limit our conclusions to STN DBS treated patients that were selected for treatment using similar exclusion criteria as those applied in this study (**i.e., the CAPSIT protocol criteria or their equivalent**, see exclusion criteria above). The lack of control group also limits application of our findings for selection purposes. Since our sample comprised of patients already selected for STN DBS treatment, the estimates could exhibit distortion due to the collider bias if generalized to a larger population of PD patients (Cinelli et al., 2022). We thus advise against using our findings as a basis for patient selection for STN DBS. Instead, practitioners should base their decision for STN DBS treatment on the current best practices (Armstrong & Okun, 2020; Defer et al., 1999) and use our findings to single out patients who could benefit from more monitoring.

Another generality constraint stems from the selection of measures used in the current study. Most importantly, there was a lack of visuo-spatial tasks in pre-surgery examination. Moreover, the cognitive outcome was evaluated by DRS-2 which although suitable for cognitive screening of global cognition does not appear to have utility in

evaluating single cognitive functions in PD (Lopez et al., 2021). **Exploiting model predictions for research decisions as described in the *Using our model as an interpretation device* section is constrained to outcomes in terms of DRS-2 scores as well. Even though in principle it is possible to convert DRS-2 scores predicted by our models to other cognitive screening tests' scales via regression equations derived from population-specific normative studies, this conversion comes with additional estimation error.**

Finally, the results of the EFA analysis can be disputed from several points of view. **As discussed above, t**he EF/Att. factor was loaded on by timed test scores and could thus be better characterized as processing speed instead. We decided to follow the naming convention established in the methods section of our article, however, the lack of time-independent executive tests constitutes a clear limitation of our data set. Moreover, several latent factors identified by the EFA were test-specific (e.g., the visuospatial memory was specific to the Family Pictures test). This issue was most pronounced in the case of verbal working memory factor which was loaded on not only by the prototypical measures of working memory capacity but also by Similarities test of WAIS-III. Some of the identified latent factors can thus represent test-specific commonalities instead of latent cognitive functions. Incidentally, these shortcomings seem to affect the EF/Att. factor the least. Notably, all three of these limitations were also observed in other recent studies that applied latent variable approach to clinically used comprehensive neuropsychological batteries in PD (Chung et al., 2021; Specketer et al., 2019). The phenomena observed in EFA results of our study may thus at least partially stem from the contemporary practice of building neuropsychological batteries according to expert consensus and warrant further investigation of a latent structure of such batteries.

Limitations and future directions

A major limitation of our study is the moderate number of missing values. To alleviate this limitation, we applied a multiple imputation technique with high number of imputations which has been shown to provide reasonable interval estimates in the

Bayesian models. However, missing data still lowered estimation precision of effects of those latent cognitive factors that were identified less consistently across imputed datasets (set shifting and spatial working memory). Another way missing data might have influenced our findings is the survivorship bias which could have led to overly optimistic estimates of the post-surgery cognitive decline rate.

Next, our results are limited to evaluating pre-surgery cognitive profile predictive of post-surgery cognitive decline. While we adopted this approach for parsimony's sake, other non-cognitive features such as demographic or clinical characteristics will likely significantly improve prediction. **Moreover, the descriptive nature of our study could be enhanced by applying effect transfer methodologies such as post-stratification to arrive at more precise prediction for other populations (Deffner et al., 2022)**

Despite these limitations, our results provide actionable information about the PD patients who are selected for STN DBS treatment based on current best practices. Based on our results, clinicians can preferentially monitor patients with a pre-surgery executive functions/attention deficit. Moreover, the cognitive profile identified in our study can serve to select within STN DBS treated patients suitable candidates for prospective clinical trials investigating effects of strategies to mitigate cognitive decline such as cognitive training, reprogramming of stimulation parameters or further DBS using a secondary target (Cappon et al., 2022). **Finally, as demonstrated above, our results can be used as a framing device for future studies investigating cognitive change in PD patients after STN DBS helping to sort out true cognitive change and measurement error.**

Conclusions

Our findings imply that STN DBS in combination with oral dopaminergic therapy is a **relatively** safe treatment option from a cognitive standpoint as it was associated with only mild annual post-surgery cognitive decline **with low probability of exceeding published reliable change cut-offs**. Pre-surgery executive functions/attention deficit appears to have a prognostic value for risk stratification with

regards to development of the post-surgery cognitive decline. Based on our models and data, we recommend considering aggregated pre-surgery results from multiple executive tests to estimate cognitive prognosis of PD patients treated with STN DBS. **We also encourage readers to use our models as interpretation devices of post-surgery screenings for cognitive changes to better understand information provided by these neuropsychological data.**

Acknowledgements

This work was supported by the Czech Ministry of Health under Grant AZV NV19-04-00233; Grant Agency of Charles University under Grant GA UK 254121; EU Joint Programme on Neurodegenerative Disease Research under Grant JPND 733051123; and by The project National Institute for Neurological Research (Programme EXCELES, ID Project No. LX22NPO5107). We further wish to thank all the patients, family members and staff from all the units that participated in the study. In particular, we wish to thank Markéta Fialová, Radka Steinbachová and Anna Rezková for their management of data collection and patient care**,** Petra Balabánová for her assistance with neuropsychological assessments**, and Martina Kvapilová for reading and commenting on later versions of this article**.

Conflict of interest

Nothing to report.

Ethical statement

The study was approved by the General University Hospital Ethics Committee in Prague, Czech Republic. All Participants provided informed consent.

Data availability

The data that support the findings of this study are not currently publicly available due institutional regulations protecting patient clinical data but are available from the corresponding author on request (may require data use agreements to be developed). The computer code used in our analysis as well as supplementary presentation of our results can be accessed at

https://github.com/josefmana/dbs_cogPRED.

References

Armstrong, M. J., & Okun, M. S. (2020). Diagnosis and Treatment of Parkinson Disease. *JAMA*, *323*(6), 548. https://doi.org/10.1001/jama.2019.22360

Barbosa, R., Guedes, L. C., Cattoni, M. B., Lobo, P. P., Caldas, A. C., Fabbri, M., Bastos, P., Valadas, A., Carvalho, H., Albuquerque, L., Reimão, S., Ferreira, A. G., Ferreira, J. J., Rosa, M. M., & Coelho, M. (2024). Long-term follow-up of subthalamic nucleus deep brain stimulation in patients with parkinson's disease: An analysis of survival and disability milestones. *Parkinsonism & Related Disorders*, *118*, 105921. https://doi.org/https://doi.org/10.1016/j.parkreldis.2023.105921

Beck, A. T., Steer, R. A., & Brown, G. (1996). *Beck depression inventory–II*. American Psychological Association (APA). https://doi.org/10.1037/t00742-000

Bezdicek, O., Lukavsky, J., Stepankova, H., Nikolai, T., Axelrod, B. N., Michalec, J., Rika, E., & Kopecek, M. (2015). The Prague Stroop Test: Normative standards in older Czech adults and discriminative validity for mild cognitive impairment in Parkinson's disease. *Journal of Clinical and Experimental Neuropsychology*, *37*(8), 794–807. https://doi.org/10.1080/13803395.2015.1057106

Bezdicek, O., Michalec, J., Nikolai, T., Havránková, P., Roth, J., Jech, R., & Rika, E. (2015). Clinical Validity of the Mattis Dementia Rating Scale in Differentiating Mild Cognitive Impairment in Parkinson's Disease and Normative Data. *Dementia and Geriatric Cognitive Disorders*, *39*(5-6), 303–311. https://doi.org/10.1159/000375365

Bezdicek, O., Motak, L., Axelrod, B. N., Preiss, M., Nikolai, T., Vyhnalek, M., Poreh, A., & Ruzicka, E. (2012). Czech Version of the Trail Making Test: Normative Data and Clinical Utility. *Archives of Clinical Neuropsychology*, *27*(8), 906–914. https://doi.org/10.1093/arclin/acs084

Bezdicek, O., Stepankova, H., Axelrod, B. N., Nikolai, T., Sulc, Z., Jech, R., Rika, E., & Kopecek, M. (2017). Clinimetric validity of the Trail Making Test Czech version in

Parkinson's disease and normative data for older adults. *The Clinical Neuropsychologist*, *31*(sup1), 42–60. https://doi.org/10.1080/13854046.2017.1324045

Bezdicek, O., Stepankova, H., Moták, L., Axelrod, B. N., Woodard, J. L., Preiss, M., Nikolai, T., Rika, E., & Poreh, A. (2014). Czech version of Rey Auditory Verbal Learning test: Normative data. *Aging, Neuropsychology, and Cognition*, *21*(6), 693–721. https://doi.org/10.1080/13825585.2013.865699

Bezdicek, O., Sulc, Z., Nikolai, T., Stepankova, H., Kopecek, M., Jech, R., & Rika, E. (2017). A parsimonious scoring and normative calculator for the Parkinson's disease mild cognitive impairment battery. *The Clinical Neuropsychologist*, *31*(6-7), 1231–1247. https://doi.org/10.1080/13854046.2017.1293161

Boel, J. A., Odekerken, V. J. J., Schmand, B. A., Geurtsen, G. J., Cath, D. C., Figee, M., van den Munckhof, P., de Haan, R. J., Schuurman, P. R., de Bie, R. M. A., Odekerken, V. J. J., Boel, J. A., van Laar, T., van Dijk, J. M. C., Mosch, A., Hoffmann, C. F. E., Nijssen, P. C. G., van Asseldonk, T., Beute, G. N., ... de Bie, R. M. A. (2016). Cognitive and psychiatric outcome 3 years after globus pallidus pars interna or subthalamic nucleus deep brain stimulation for parkinson's disease. *Parkinsonism & Related Disorders*, *33*, 90–95. https://doi.org/https://doi.org/10.1016/j.parkreldis.2016.09.018

Bove, F., Fraix, V., Cavallieri, F., Schmitt, E., Lhommée, E., Bichon, A., Meoni, S., Pélissier, P., Kistner, A., Chevrier, E., Ardouin, C., Limousin, P., Krack, P., Benabid, A. L., Chabardès, S., Seigneuret, E., Castrioto, A., & Moro, E. (2020). Dementia and subthalamic deep brain stimulation in Parkinson disease. *Neurology*, *95*(4). https://doi.org/10.1212/wnl.0000000000009822

Bratsos, S. P., Karponis, D., & Saleh, S. N. (2018). Efficacy and Safety of Deep Brain Stimulation in the Treatment of Parkinson's Disease: A Systematic Review and Meta-analysis of Randomized Controlled Trials. *Cureus*. https://doi.org/10.7759/cureus.3474

Browne, M. W., & Cudeck, R. (1992). Alternative Ways of Assessing Model Fit.

Sociological Methods & Research, *21*(2), 230–258. https://doi.org/10.1177/0049124192021002005

Burgess, P. W. (2014). Theory and Methodology in Executive Function Research. In P. Rabbitt (Ed.), *Methodology of Frontal and Executive Function* (pp. 87–121). Psychology Press.

Bürkner, P.-C. (2017). **brms**: An *R* Package for Bayesian Multilevel Models Using *Stan*. *Journal of Statistical Software*, *80*(1). https://doi.org/10.18637/jss.v080.i01

Bürkner, P.-C., Gabry, J., & Vehtari, A. (2020). Efficient leave-one-out cross-validation for Bayesian non-factorized normal and Student-t models. *Computational Statistics*, *36*(2), 1243–1261. https://doi.org/10.1007/s00180-020-01045-4

Cappon, D., Gratwicke, J., Zrinzo, L., Akram, H., Hyam, J., Hariz, M., Limousin, P., Foltynie, T., & Jahanshahi, M. (2022). Deep Brain Stimulation of the Nucleus Basalis of Meynert for Parkinson's Disease Dementia: A 36 Months Follow Up Study. *Movement Disorders Clinical Practice*, *9*(6), 765–774. https://doi.org/10.1002/mdc3.13510

Castrioto, A., Debû, B., Cousin, E., Pelissier, P., Lhommée, E., Bichon, A., Schmitt, E., Kistner, A., Meoni, S., Seigneuret, E., Chabardes, S., Krack, P., Moro, E., & Fraix, V. (2022). Long-term independence and quality of life after subthalamic stimulation in Parkinson disease. *European Journal of Neurology*, *29*(9), 2645–2653. https://doi.org/10.1111/ene.15436

Chung, S. J., Yoo, H. S., Lee, H. S., Lee, Y. H., Baik, K., Jung, J. H., Ye, B. S., Sohn, Y. H., & Lee, P. H. (2021). Baseline cognitive profile is closely associated with long-term motor prognosis in newly diagnosed Parkinson's disease. *Journal of Neurology*, *268*(11), 4203–4212. https://doi.org/10.1007/s00415-021-10529-2

Ciharova, M., Cígler, H., Dostálová, V., ivicová, G., & Bezdicek, O. (2020). Beck depression inventory, second edition, Czech version: demographic correlates, factor structure and comparison with foreign data. *International Journal of Psychiatry in Clinical Practice*, *24*(4), 371–379.

https://doi.org/10.1080/13651501.2020.1775854

Cinelli, C., Forney, A., & Pearl, J. (2022). A Crash Course in Good and Bad Controls. *Sociological Methods & Research*, 004912412210995. https://doi.org/10.1177/00491241221099552

Combs, H. L., Folley, B. S., Berry, D. T. R., Segerstrom, S. C., Han, D. Y., Anderson-Mooney, A. J., Walls, B. D., & Horne, C. van. (2015). Cognition and Depression Following Deep Brain Stimulation of the Subthalamic Nucleus and Globus Pallidus Pars Internus in Parkinson's Disease: A Meta-Analysis. *Neuropsychology Review*, *25*(4), 439–454. https://doi.org/10.1007/s11065-015-9302-0

David, F. J., Munoz, M. J., & Corcos, D. M. (2020). The effect of STN DBS on modulating brain oscillations: Consequences for motor and cognitive behavior. *Experimental Brain Research*, *238*(7-8), 1659—1676. https://doi.org/10.1007/s00221-020-05834-7

Defer, G.-L., Widner, H., Marié, R.-M., Rémy, P., & Levivier, M. (1999). Core assessment program for surgical interventional therapies in parkinson's disease (CAPSIT-PD). *Movement Disorders*, *14*(4), 572–584. https://doi.org/https://doi.org/10.1002/1531-8257(199907)14:4%3C572::AID-MDS1005%3E3.0.CO;2-C

Deffner, D., Rohrer, J. M., & McElreath, R. (2022). A causal framework for cross-cultural generalizability. *Advances in Methods and Practices in Psychological Science*, *5*(3), 25152459221106366. https://doi.org/10.1177/25152459221106366

Ewert, S., Plettig, P., Li, N., Chakravarty, M. M., Collins, D. L., Herrington, T. M., Kühn, A. A., & Horn, A. (2018). Toward defining deep brain stimulation targets in MNI space: A subcortical atlas based on multimodal MRI, histology and structural connectivity. *NeuroImage*, *170*, 271–282. https://doi.org/https://doi.org/10.1016/j.neuroimage.2017.05.015

Frydrychová, Z., Kopeek, M., Bezdicek, O., & Georgi Stepankova, H. (2018). Czech

normative study of the Revised Rey Auditory Verbal Learning Test (RAVLT) in older adults. *Ceskoslovenska Psychologie*, *62*(4), 330–349.

Gelman, A., Hill, J., & Yajima, M. (2012). Why We (Usually) Don't Have to Worry About Multiple Comparisons. *Journal of Research on Educational Effectiveness*, *5*(2), 189–211. https://doi.org/10.1080/19345747.2011.618213

Gelman, A., & Vákár, M. (2021). Slamming the sham: A bayesian model for adaptive adjustment with noisy control data. *Statistics in Medicine*, *40*(15), 3403–3424. https://doi.org/https://doi.org/10.1002/sim.8973

Gruber, D., Calmbach, L., Kühn, A. A., Krause, P., Kopp, U. A., Schneider, G.-H., & Kupsch, A. (2019). Longterm outcome of cognition, affective state, and quality of life following subthalamic deep brain stimulation in Parkinson's disease. *Journal of Neural Transmission*, *126*(3), 309–318. https://doi.org/10.1007/s00702-019-01972-7

Harman, H. H., & Jones, W. H. (1966). Factor analysis by minimizing residuals (minres). *Psychometrica*, *31*(3), 351–368.

Hentz, J. G., Mehta, S. H., Shill, H. A., Driver-Dunckley, E., Beach, T. G., & Adler, C. H. (2015). Simplified conversion method for unified Parkinson's disease rating scale motor examinations. *Movement Disorders*, *30*(14), 1967–1970. https://doi.org/10.1002/mds.26435

Horn, A., & Kühn, A. A. (2015). Lead-DBS: A toolbox for deep brain stimulation electrode localizations and visualizations. *NeuroImage*, *107*, 127–135. https://doi.org/https://doi.org/10.1016/j.neuroimage.2014.12.002

Horn, A., Li, N., Dembek, T. A., Kappel, A., Boulay, C., Ewert, S., Tietze, A., Husch, A., Perera, T., Neumann, W.-J., Reisert, M., Si, H., Oostenveld, R., Rorden, C., Yeh, F.-C., Fang, Q., Herrington, T. M., Vorwerk, J., & Kühn, A. A. (2019). Lead-DBS v2: Towards a comprehensive pipeline for deep brain stimulation imaging. *NeuroImage*, *184*, 293–316. https://doi.org/https://doi.org/10.1016/j.neuroimage.2018.08.068

Hughes, A. J., Daniel, S. E., Kilford, L., & Lees, A. J. (1992). Accuracy of clinical

diagnosis of idiopathic Parkinson's disease: a clinico-pathological study of 100 cases. *Journal of Neurology, Neurosurgery & Psychiatry*, *55*(3), 181–184. https://doi.org/10.1136/jnnp.55.3.181

Iannone, R. (2022). *DiagrammeR: Graph/network visualization.* https://CRAN.R-project.org/package=DiagrammeR

Jahanshahi, M., Leimbach, F., & Rawji, V. (2022). Short and long-term cognitive effects of subthalamic deep brain stimulation in parkinson's disease and identification of relevant factors. *Journal of Parkinson's Disease*, *12*(7), 2191—2209. https://doi.org/10.3233/jpd-223446

Jech, R., Mueller, K., Urgosík, D., Sieger, T., Holiga, ., Rika, F., Duek, P., Havránková, P., Vymazal, J., & Rika, E. (2012). The Subthalamic Microlesion Story in Parkinson's Disease: Electrode Insertion-Related Motor Improvement with Relative Cortico-Subcortical Hypoactivation in fMRI. *PLoS ONE*, *7*(11), e49056. https://doi.org/10.1371/journal.pone.0049056

Jech, R., Ruzicka, E., Ugosik, D., Serranova, T., Volfova, M., Novakova, O., Roth, J., Dusek, P., & Mecir, P. (2006). Deep brain stimulation of the subthalamic nucleus affects resting EEG and visual evoked potentials in Parkinson's disease. *Clinical Neurophysiology*, *117*(5), 1017–1028. https://doi.org/10.1016/j.clinph.2006.01.009

Josse, J., & Husson, F. (2016). *{missMDA}: A package for handling missing values in multivariate data analysis*. *70*. https://doi.org/10.18637/jss.v070.i01

Jurica, P. J., Leitten, C. L., & Mattis, S. (2001). *Dementia rating scale-2 (DRS-2) professional manual*. Psychological Assessment Resources.

Kim, H.-J., Jeon, B. S., Paek, S. H., Lee, K.-M., Kim, J.-Y., Lee, J.-Y., Kim, H. J., Yun, J. Y., Kim, Y. E., Yang, H.-J., & Ehm, G. (2014). Long-term cognitive outcome of bilateral subthalamic deep brain stimulation in Parkinson's disease. *Journal of Neurology*, *261*(6), 1090–1096. https://doi.org/10.1007/s00415-014-7321-z

Kishore, A., Krishnan, S., Pisharady, K., Rajan, R., Sarma, S., & Sarma, P. (2019). Predictors of dementia-free survival after bilateral subthalamic deep brain

stimulation for Parkinson's disease. *Neurology India*, *67*(2), 459. https://doi.org/10.4103/0028-3886.258056

Litvan, I., Goldman, J. G., Tröster, A. I., Schmand, B. A., Weintraub, D., Petersen, R. C., Mollenhauer, B., Adler, C. H., Marder, K., Williams-Gray, C. H., Aarsland, D., Kulisevsky, J., Rodriguez-Oroz, M. C., Burn, D. J., Barker, R. A., & Emre, M. (2012). Diagnostic criteria for mild cognitive impairment in Parkinson's disease: *Movement* Disorder Society Task Force guidelines. *Movement Disorders*, *27*(3), 349–356. https://doi.org/10.1002/mds.24893

Lofredi, R., Auernig, C.-G., Ewert, S., Irmen, F., Steiner, L. A., Scheller, U., van Wijk, B. C. M., Oxenford, S., Kühn, A. A., & Horn, A. (2022). Interrater reliability of deep brain stimulation electrode localizations. *NeuroImage*, *262*, 119552. https://doi.org/https://doi.org/10.1016/j.neuroimage.2022.119552

Lopez, F. V., Kenney, L. E., Ratajska, A., Jacobson, C. E., & Bowers, D. (2021). What does the Dementia Rating Scale-2 measure? The relationship of neuropsychological measures to DRS-2 total and subscale scores in non-demented individuals with Parkinson's disease. *The Clinical Neuropsychologist*, *37*(1), 174–193. https://doi.org/10.1080/13854046.2021.1999505

Lundberg, I., Johnson, R., & Stewart, B. M. (2021). What Is Your Estimand? Defining the Target Quantity Connects Statistical Evidence to Theory. *American Sociological Review*, *86*(3), 532–565. https://doi.org/10.1177/00031224211004187

Mazancova, A. F., Rika, E., Jech, R., & Bezdicek, O. (2020). Test the Best: Classification Accuracies of Four Cognitive Rating Scales for Parkinson's Disease Mild Cognitive Impairment. *Archives of Clinical Neuropsychology*, *35*(7), 1069–1077. https://doi.org/10.1093/arclin/acaa039

Mehanna, R., Bajwa, J. A., Fernandez, H., & Wagle Shukla, A. A. (2017). Cognitive Impact of Deep Brain Stimulation on Parkinson's Disease Patients. *Parkinson's Disease*, *2017*, 1–15. https://doi.org/10.1155/2017/3085140

Michalec, J., Bezdicek, O., Nikolai, T., Harsa, P., Jech, R., Silhan, P., Hyza, M., Ruzicka, E., & Shallice, T. (2017). A Comparative Study of Tower of London

Scoring Systems and Normative Data. *Archives of Clinical Neuropsychology*. https://doi.org/10.1093/arclin/acw111

Nikolai, T., Stepankova, H., Michalec, J., Bezdicek, O., Horáková, K., Marková, H., Ruzicka, E., & Kopecek, M. (2015). Tests of verbal fluency, czech normative study in older patients. *eská a Slovenská Neurologie a Neurochirurgie*, *78/111*(3), 292–299. https://doi.org/10.14735/amcsnn2015292

Park, T., & Casella, G. (2008). The Bayesian Lasso. *Journal of the American Statistical Association*, *103*(482), 681–686. https://doi.org/10.1198/016214508000000337

Parsons, T. D., Rogers, S. A., Braaten, A. J., Woods, S. P., & Tröster, A. I. (2006). Cognitive sequelae of subthalamic nucleus deep brain stimulation in Parkinson's disease: a meta-analysis. *The Lancet Neurology*, *5*(7), 578–588. https://doi.org/10.1016/s1474-4422(06)70475-6

Partington, J. E., & Leiter, R. G. (1949). Partington's Pathways Test. *Psychological Service Center Journal*, *1*, 11–20.

Pedersen, T. L. (2020). *Patchwork: The composer of plots*. https://CRAN.R-project.org/package=patchwork

Pedraza, O., Smith, G. E., Ivnik, R. J., Willis, F. B., Ferman, T. J., Petersen, R. C., Graff-Radford, N. R., & Lucas, J. A. (2007). Reliable change on the dementia rating scale. *Journal of the International Neuropsychological Society*, *13*(4), 716–720. https://doi.org/10.1017/S1355617707070920

R Core Team. (2024). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Reich, M. M., Hsu, J., Ferguson, M., Schaper, F. L. W. V. J., Joutsa, J., Roothans, J., Nickl, R. C., Frankemolle-Gilbert, A., Alberts, J., Volkmann, J., & Fox, M. D. (2022). A brain network for deep brain stimulation induced cognitive decline in Parkinson's disease. *Brain*, *145*(4), 1410–1421. https://doi.org/10.1093/brain/awac012

Revelle, W. (2022). *Psych: Procedures for psychological, psychometric, and personality research*. https://CRAN.R-project.org/package=psych

Schupbach, W. M. M. (2005). Stimulation of the subthalamic nucleus in Parkinson's disease: a 5 year follow up. *Journal of Neurology, Neurosurgery & Psychiatry*, *76*(12), 1640–1644. https://doi.org/10.1136/jnnp.2005.063206

Schüpbach, W. M. M., Maltête, D., Houeto, J. L., Montcel, S. T. du, Mallet, L., Welter, M. L., Gargiulo, M., Béhar, C., Bonnet, A. M., Czernecki, V., Pidoux, B., Navarro, S., Dormont, D., Cornu, P., & Agid, Y. (2007). Neurosurgery at an earlier stage of parkinson disease. *Neurology*, *68*(4), 267–271. https://doi.org/10.1212/01.wnl.0000250253.03919.fb

Shallice, T. (1982). Specific impairments of planning. *Philosophical Transactions of the Royal Society of London. B, Biological Sciences*, *298*(1089), 199–209. https://doi.org/10.1098/rstb.1982.0082

Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis*. Oxford University PressNew York. https://doi.org/10.1093/acprof:oso/9780195152968.001.0001

Smeding, H. M. M., Speelman, J. D., Huizenga, H. M., Schuurman, P. R., & Schmand, B. (2009). Predictors of cognitive and psychosocial outcome after STN DBS in Parkinson's Disease. *Journal of Neurology, Neurosurgery & Psychiatry*, *82*(7), 754–760. https://doi.org/10.1136/jnnp.2007.140012

Specketer, K., Zabetian, C. P., Edwards, K. L., Tian, L., Quinn, J. F., Peterson-Hiller, A. L., Chung, K. A., Hu, S.-C., Montine, T. J., & Cholerton, B. A. (2019). Visuospatial functioning is associated with sleep disturbance and hallucinations in nondemented patients with Parkinson's disease. *Journal of Clinical and Experimental Neuropsychology*, *41*(8), 803–813. https://doi.org/10.1080/13803395.2019.1623180

Spielberger, C. D., Gorsuch, R. L., Lushene, R., Vagg, P. R., & Jacobs, G. A. (1983). *Manual for the state-trait anxiety inventory*. Palo Alto, CA: Consulting Psychologists Press.

Stan Development Team. (2020). *Stan modeling language users guide and reference manual, version 2.21.0*. http://mc-stan.org/

Thomson, G. (1951). *The factorial analysis of human ability*. University of London Press.

Tomlinson, C. L., Stowe, R., Patel, S., Rick, C., Gray, R., & Clarke, C. E. (2010). Systematic review of levodopa dose equivalency reporting in Parkinson's disease. *Movement Disorders*, *25*(15), 2649–2653. https://doi.org/10.1002/mds.23429

Tuerlinckx, F., Rijmen, F., Verbeke, G., & De Boeck, P. (2006). Statistical inference in generalized linear mixed models: A review. *British Journal of Mathematical and Statistical Psychology*, *59*(2), 225–255. https://doi.org/10.1348/000711005x79857

Urgosik, D., Jech, R., Ruzicka, E., Ruzicka, F., Liscák, R., & Vladyka, V. (2011). Deep brain stimulation in movement disorders: a Prague-center experience. *Casopis Lekaru Ceskych*, *150*(4-5), 223–228.

Van Bork, R., Rhemtulla, M., Sijtsma, K., & Borsboom, D. (2023). A causal theory of error scores. *Psychological Methods*. https://doi.org/10.1037/met0000521

Vehtari, A., Simpson, D., Gelman, A., Yao, Y., & Gabry, J. (2015). *Pareto smoothed importance sampling*. https://doi.org/10.48550/ARXIV.1507.02646

Wechsler, D. (2010). *Wechsler adult intelligence scale - third revision*. Hogrefe - Testcentrum.

Wechsler, D. (2011). *Wechsler memory scale -third edition abbreviated*. Hogrefe - Testcentrum.

Whitney, P., & Hinson, J. M. (2010). *Measurement of cognition in studies of sleep deprivation* (pp. 37–48). Elsevier. https://doi.org/10.1016/b978-0-444-53702-7.00003-8

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. https://ggplot2.tidyverse.org

Wilke, C. O. (2024). *Ggridges: Ridgeline plots in 'ggplot2'*. https://wilkelab.org/ggridges/

Wood, S. N., Scheipl, F., & Faraway, J. J. (2012). Straightforward intermediate rank tensor product smoothing in mixed models. *Statistics and Computing*, *23*(3),

341–360. https://doi.org/10.1007/s11222-012-9314-z

Xu, Y., Qin, G., Tan, B., Fan, S., An, Q., Gao, Y., Fan, H., Xie, H., Wu, D., Liu, H., Yang, G., Fang, H., Xiao, Z., Zhang, J., Zhang, H., Shi, L., & Yang, A. (2023). Deep brain stimulation electrode reconstruction: Comparison between lead-DBS and surgical planning system. *Journal of Clinical Medicine*, *12*(5). https://doi.org/10.3390/jcm12051781

Yarkoni, T. (2020). The generalizability crisis. *Behavioral and Brain Sciences*, *45*. https://doi.org/10.1017/s0140525x20001685

Zwet, E. W. van. (2019). A default prior for regression coefficients. *Statistical Methods in Medical Research*, *28*(12), 3799–3807. https://doi.org/10.1177/0962280218817792

Zwet, E. W. van, & Cator, E. A. (2021). The significance filter, the winner's curse and the need to shrink. *Statistica Neerlandica*, *75*(4), 437–452. https://doi.org/https://doi.org/10.1111/stan.12241

Figure 1

Patients inclusion/exclusion flowchart. The plot shows reasons for excluding patients from the total sample of 200 patients considered for subthalamic nucleus deep brain stimulation treatment.

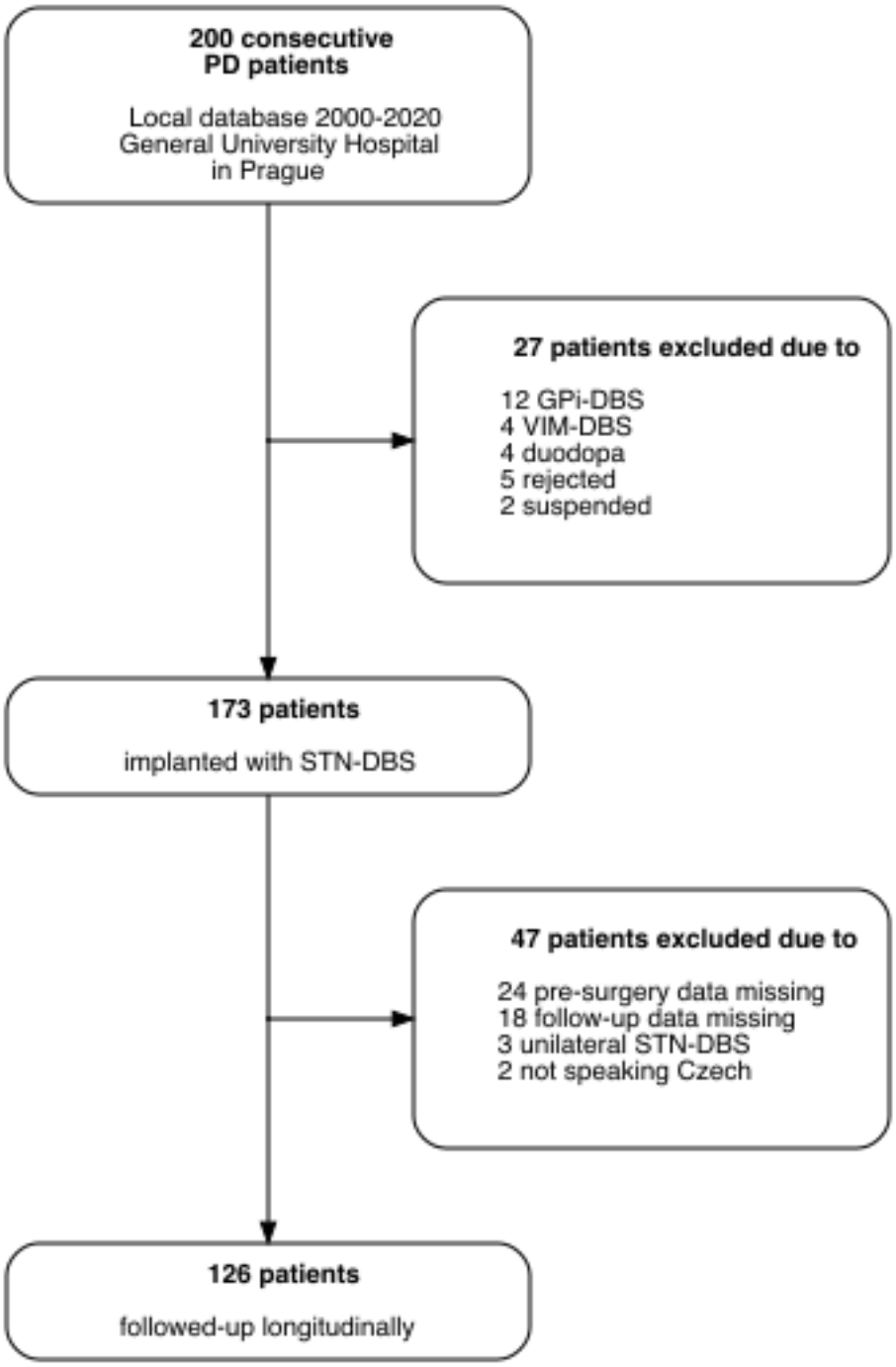

Figure 2

Distribution of assessments. Distribution of (A) follow-up years and (B) number of assessments per patient for N = 126 patients. Negative values on horizontal axis in (A) represent pre-surgery assessments, number of assessments in (B) includes one pre-surgery and various number of post-surgery assessments.

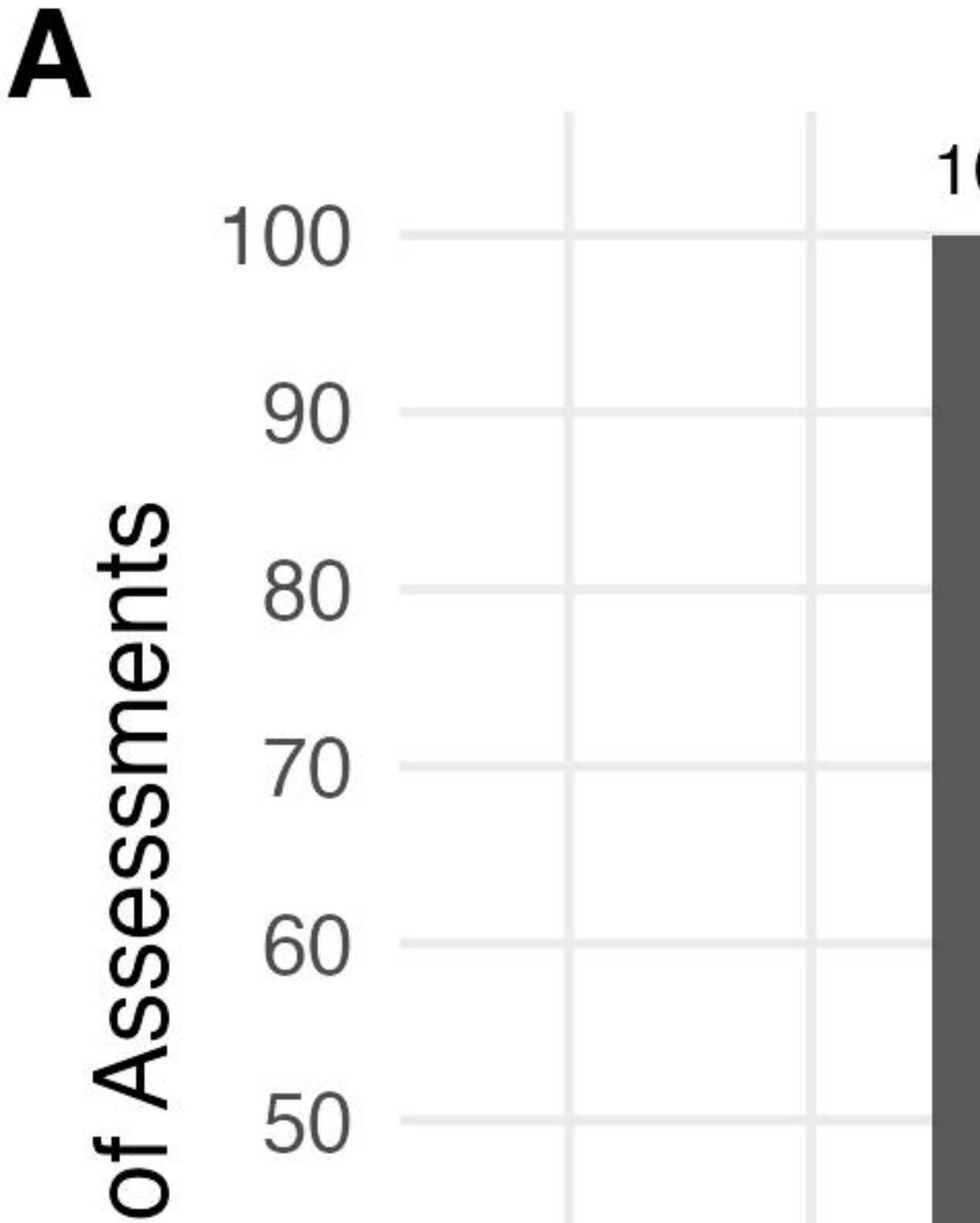

Figure 3

Comparison of linear versus non-linear models of the longitudinal cognitive trajectory. The figure shows expected true scores in Mattis Dementia Rating Scale (DRS-2) as estimated by the descriptive linear (pink line) and non-linear (black line) models with their 95% posterior probability intervals (shades).

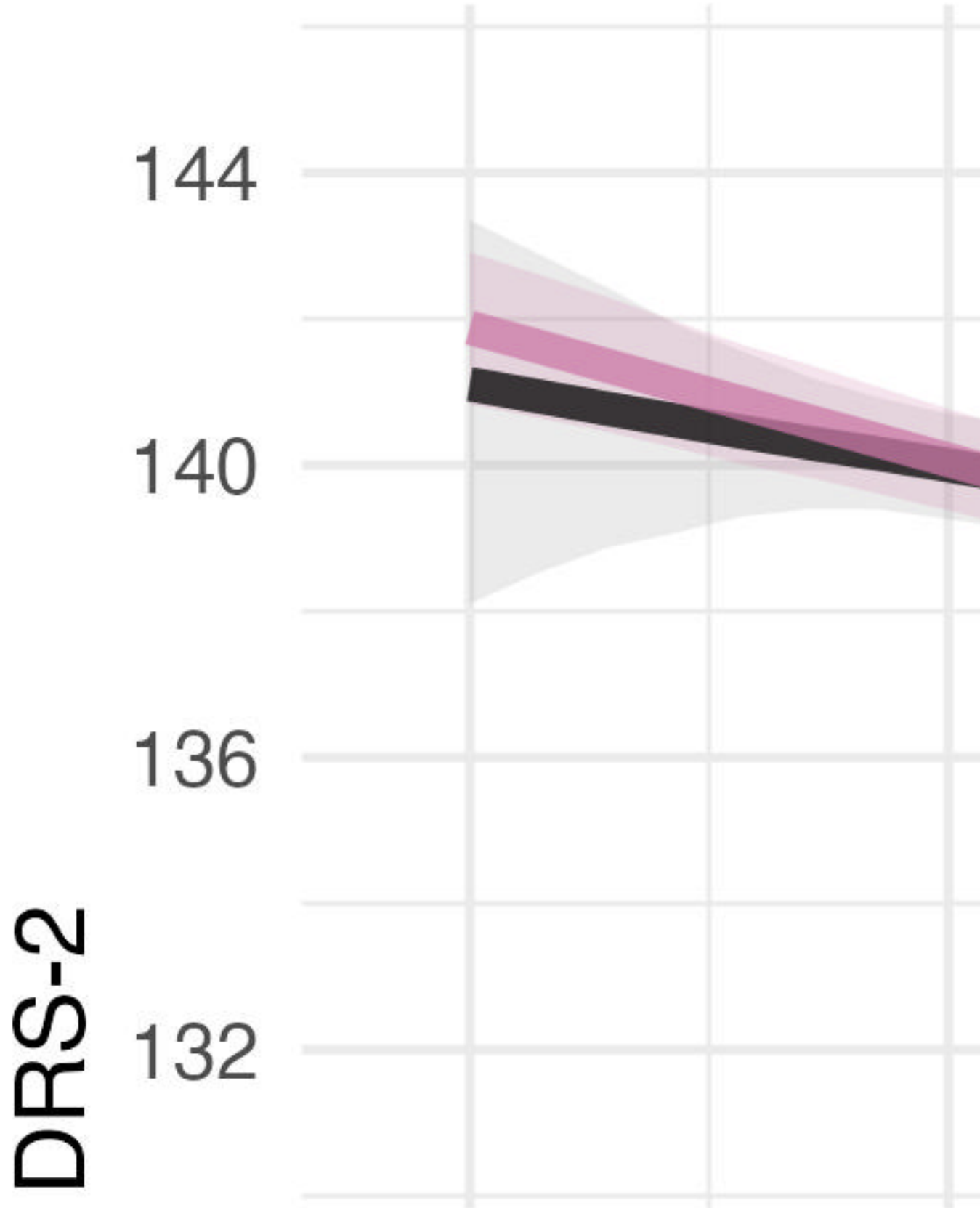

Figure 4

Interaction terms of the "test scores" (left column) and the "factor scores" (right column) models predicting post-surgery cognitive decline. Each density represents full posterior estimate of additive predictive value of the cognitive variable (test or factor) listed on the ordinate. All cognitive predictors were scaled such that negative values mean negative effect of pre-surgery deficit on prediction of longitudinal cognitive trajectory. Vertical lines represent zero (no effect, solid black line) and average post-surgery decline in our sample accroding to the descriptive linear model predictive before (blue dashed line). Acronyms are explained in the text.