Evaluation of Diagnostic Concordance Between Algorithms for Parkinson's Disease Dementia

Martina Mana, Josef Mana, Tereza Uhrova, Robert Jech, and Ondrej Bezdicek

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

Author Note

Martina Mana https://orcid.org/0009-0007-4665-3946

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

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

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

Supported by The project National Institute for Neurological Research (Programme EXCELES, ID Project No. LX22NPO5107) - Funded by the European Union – Next Generation EU; Charles University: Cooperatio Program in Neuroscience; General University Hospital in Prague project MH CZ-DRO-VFN64165. Author roles were classified using the Contributor Role Taxonomy (CRediT; https://credit.niso.org/) as follows: Martina Mana: Conceptualization, Data curation, Writing - original draft; Josef Mana: Conceptualization, Data curation, Investigation, Formal analysis, Software, Methodology, Project administration, Validation, Writing - original draft; Tereza Uhrova: Investigation; Robert Jech: Funding acquisition, Resources, Writing - review & editing; Ondrej Bezdicek: Investigation, Data curation, Funding acquisition, Conceptualization, Project administration, Supervision, Writing - original draft

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

Abstract

Background: A significant number of patients with Parkinson’s disease (PD) gradually progress to Parkinson’s disease dementia (PDD). A recent “Call for Change” (Kulisevsky et al., 2024) proposes updates to the current diagnostic criteria for PDD, including the use of different screening tests and broader functional assessments. Objective: We aimed to evaluate the diagnostic concordance between several algorithms for PDD based on Level I (i.e., screening) criteria and to assess their predictive validity for Level II (i.e., neuropsychological battery) diagnosis. Methods: We conducted a cross-sectional retrospective analysis of 204 patients diagnosed with PD. All patients underwent a comprehensive neuropsychological battery. A total of 68 diagnostic algorithms were operationalised based on the combinations of various cognitive assessments, including the Mini Mental State Examination (MMSE), the Montreal Cognitive Assessment (MoCA), and the shortened version of MoCA (sMoCA). Functional impairment was based on the Functional Assessment Questionnaire (FAQ), specifically FAQ item 9 or the FAQ total score. Concordance between algorithms was assessed using pairwise Cohen’s . The predictive validity of Level I algorithms for Level II diagnosis was assessed using Exact Binomial Tests. Results: PDD rate estimates varied widely across algorithms from 2.00% up to 16.75%. Higher rates were associated with functional impairment based on the definition of FAQ score (= 0.75, = 0.86). Moderate to high concordance was observed among algorithms using the same functional impairment definition. MoCA-based algorithms most accurately predicted Level II classification, especially those using Clock Drawing to assess executive function. Conclusions: Diagnostic outcomes for PDD are sensitive to the choice of cognitive and functional instruments. Our results support some of the proposed changes to PDD diagnostic criteria on Level I, emphasising the consistency of MoCA-based assessments and comprehensive functional impairment evaluation.

Keywords: Parkinson’s disease, dementia, diagnostic criteria, functional assessment, cognitive assessment

Evaluation of Diagnostic Concordance Between Algorithms for Parkinson's Disease Dementia

Introduction

Parkinson's disease (PD) is a neurodegenerative disorder typically characterized by a progressive onset of motor symptoms, including rigidity, bradykinesia, postural instability and resting tremor. Moreover, patients suffer from a range of non-motor impairments (Postuma et al., 2015), particularly cognitive decline. This factor might result in Parkinson's disease dementia (PDD) in a subset of patients (Meireles & Massano, 2012).

According to a recent meta-analysis, approximately one-quarter of PD patients is likely to be diagnosed with PDD (Sousa et al., 2022). However, reported PDD rate estimates vary widely, ranging from 14% up to 55%, depending on the methodological criteria employed (Sousa et al., 2022). Moreover, factors such as patients' sex (Cereda et al., 2016), age and disease duration appear to modulate the risk of cognitive decline and PDD (Oh et al., 2016; Rana et al., 2011).

Despite the clinical relevance of PDD, its diagnosis remains complex. A milestone in research of PDD was the publication of diagnostic criteria established in 2007 by the International Parkinson and Movement Disorder Society (MDS) (Dubois et al., 2007). In these criteria, the MDS introduced a two-levelled system for PDD detection. Level I consists of brief cognitive assessments, while Level II involves comprehensive neuropsychological testing across cognitive domains (Emre et al., 2007).

The original Level I algorithm included eight conditions that had to be satisfied simultaneously in order to diagnose probable PDD. These included: 1) diagnosis of PD proposed by the Queen Square Brain Bank; 2) PD onset prior to the PDD emergence; 3) evidence of global cognitive impairment (MMSE score < 26 points); 4) cognitive deficit interference with the IADL (assessed by the pill questionnaire or caregiver interview); 5) impairment in at least two cognitive domains, namely memory, attention, visuo-constructive abilities and executive function; 6) there was absence of Major Depressive Disorder; 7) absence of delirium; and 8) exclusion of other abnormalities and potential causes of dementia (Dubois et al., 2007).

Currently, efforts are focused on refining this PDD diagnostic framework. A recent call for a change pinpoints limitations regarding the original criteria and suggest various updates to enhance their utility (Kulisevsky et al., 2024). Proposed suggestions include replacement of the Mini Mental State Examination (MMSE) by the Montreal Cognitive Assessment (MoCA), which is more sensitive to PD specific cognitive impairment; expansion of instrumental activities of daily living (IADL) evaluation; inclusion of language assessment; recognition of anxiety as one **of** the neuropsychiatric symptoms relevant in PDD; and integration of biomarkers.

In light of these proposals, the current study aims to evaluate the diagnostic concordance between the original MDS Level I PDD criteria (Dubois et al., 2007; Emre et al., 2007) and a modified framework based on the recent call for change (Kulisevsky et al., 2024). Furthermore, both Level I diagnostic approaches are compared to PDD diagnosed on Level II. The study aims to address the following research objectives (RO): (RO1) To estimate the PDD rate and evaluate the diagnostic variability and concordance across different PDD criteria. (RO2) To identify specific diagnostic components contributing to PDD classification variability across the applied criteria.

Methods

Participants

This study retrospectively analyzed clinical data from a cohort of patients with PD at the General University Hospital in Prague. All patients were diagnosed with idiopathic PD by a movement disorder specialist according to the MDS Clinical Diagnostic Criteria for PD (Postuma et al., 2015). Clinical records spanning August 2014 to February 2025 were examined. All participants were candidates for Deep Brain Stimulation (DBS) treatment and underwent neuropsychological evaluation conducted by a trained clinical psychologist (OB) as part of standard preoperative assessments for DBS eligibility at the General University Hospital in Prague.

Neuropsychological Assessment

Cognitive performance was evaluated at both Level I and Level II according to the standard MDS battery for Parkinson's Disease Mild Cognitive Impairment (PD-MCI) Bezdicek et al. (2017). Cognitive performance at Level I was assessed by **the MMSE (Folstein et al., 1975; Stepankova et al., 2015) and the MoCA** (Kopecek et al., 2016; Nasreddine et al., 2005). The comprehensive neuropsychological assessment at Level II evaluated five cognitive domains through specific tests: attention and working memory assessed by Trail Making Test Part A (TMT-A) (Bezdicek et al., 2012; Reitan, 2004), and WAIS Digit Span Backward (WAIS DSB) (Wechsler, 1997), executive function by Categorical Verbal Fluency (CF) (Nikolai et al., 2015), and subtest from the Prague Stroop Test – Colors (PST-C) (Bezdicek et al., 2021), language by the WAIS Similarities subtest (Wechsler, 1997), and the Boston Naming Test (BNT-60) (Kaplan et al., 1983; Zemanová et al., 2016), memory by the Rey Auditory Verbal Learning Test (RAVLT) (Bezdicek et al., 2013; Frydrychová et al., 2018; Rey, 1964) delayed recall, and the Brief Visual Memory Test–Revised (BVMTR) (Benedict, 1997; Havlík et al., 2020) delayed recall, or WAIS Family Pictures subtest (Wechsler, 1997) delayed recall, visuospatial function assessed by the Judgment of Line Orientation Test (JoLO) (Benton et al., 1983), and Clock Drawing Test (CLOX) (Royall et al., 1998).

The Functional Activities Questionnaire (FAQ) (Bezdicek, Nikolai, et al., 2016; Pfeffer et al., 1982) was administered to assess functional impairment. The Beck Depression Inventory-II (BDI-II) (Beck et al., 1996; Ciharova et al., 2020) and State-Trait Anxiety Inventory (STAI) (Mullner et al., 1980; Spielberger et al., 1983) were used to assess neuropsychiatric status.

Diagnostic algorithms for probable Parkinson’s Disease Dementia

In this study, we applied three distinct sets of diagnostic algorithms for probable PDD at Level I. The first set was based on the original framework (Dubois et al., 2007), which utilized the Mini-Mental State Examination (MMSE) as a global cognitive screening tool, supplemented by assessments of attention, executive function, visuospatial abilities, and memory. The second set of algorithms was based on the recent call for change of dementia diagnostic guidelines (Kulisevsky et al., 2024), which advocates for more sensitive cognitive domain assessments in the context of PD. This updated approach incorporated specific items from the MoCA. The third approach applied the Czech version of the shortened Montreal Cognitive Assessment (sMoCA) (Bezdicek et al., 2020), a time-efficient modification designed to measure global cognitive performance using a reduced testing protocol that omits items providing redundant information. **The sMoCA has been validated in Czech PD cohort (Bezdicek et al., 2020) and shown to be sensitive to cognitive deficits while lowering patient burden (Roalf et al., 2017). We included the sMoCA in our study for its clinical utility in pre-surgical settings, where time restrictions and patients' fatigue often limit the feasibility of longer assessments. Moreover, the Czech validation study reported comparable diagnostic accuracy between MoCA (AUC = 0.815) and sMoCA (AUC = 0.796) for distinguishing PD-MCI from PD-NC, supporting the sMoCA as a suitable and efficient alternative.**

Lastly, the fourth approach followed the Level II protocol for diagnosis of PDD and Mild Cognitive Impairment in PD (PD-MCI) (Dubois et al., 2007; Litvan et al., 2012). The Level II methodology, including the use of a regression-based normative scoring approach, has been detailed in a prior study (Bezdicek et al., 2017). In this study, the thresholds for cognitive impairment at Level II were set at $z \leq -1.5$. All non-cognitive criteria of probable PDD (i.e., diagnosis of PD that developed before dementia and absence of Major Depression, delirium or other abnormalities that obscure diagnosis) held true for all patients in the sample according to the psychiatric and neurological examinations.

For each of these diagnostic approaches, we applied two operationalizations of deficits in Instrumental Activities of Daily Living (IADL). First, we utilized FAQ item 9, which approximates the pill questionnaire from the original criteria (Dubois et al., 2007) employing a cut-off score of 2 points or higher. Second, we applied the entire Functional Activities Questionnaire (FAQ) as suggested in the call for change (Kulisevsky et al., 2024), employing a cut-off score of 7 points or higher based on Czech normative data (Bezdicek et al., 2011). These methodologies resulted in a total of 68 algorithms, which were distributed across different diagnostic criteria: 4 MMSE-based, 60 MoCA-based, 2 sMoCA-based, and 2 based on the Level II battery (see Figure 1, Table 1 and Appendix Table A1 for the exact specification of each algorithm).

Finally, all patients were systematically evaluated for the presence of neuropsychiatric symptoms, including depression, apathy, anxiety, psychosis, and delirium, by a trained neuropsychiatrist (TU) experienced in the assessment of patients with movement disorders. Because severe psychiatric symptoms form exclusion criteria for the diagnosis of probable PDD (conditions 6–8; p. 1), all patients classified as PDD were double-checked in hospital records to confirm the absence of such confounding symptoms.

[Insert Table 1 here]

Statistical Analyses

Following the framework proposed Lundberg et al. (2021), in this study we explicitly connect our research objectives and their corresponding theoretical (i.e., targets of inference) and empirical (i.e., data-driven) estimands to statistical estimates. The theoretical estimand refers to a unit-specific quantity defined over a target population and represents the ideal quantity that would address the research question under optimal conditions, such as access to complete population data or perfect experimental control. In contrast, the empirical estimand corresponds to the quantity that is actually computable using the available dataset, given real world constraints. A full description of the study's estimands and their relation to our research objectives is presented in the Appendix (see Table A2).

To address study objectives, we started by repeatedly assigning each patient the diagnosis of probable PDD based on each PDD algorithm listed in Table 1 (see also Table A1) resulting in a 204 (patients) × 68 (algorithms) matrix where each cell indicates whether a patient (row) meets criteria for probable PDD according to an algorithm (column). PDD rate estimates were computed as $\frac{N_{PDD}}{N_{total}}$ separately for each algorithm. The predictive value of age and sex was then evaluated by fitting a set of logistic regressions, one for each algorithm for probable PDD, whereby the probable PDD was predicted by age, sex and their interaction.

Next, a set of two class cross-tabulations with associated statistics was computed for each pair of algorithms via the `confusionMatrix()` function from the R package *caret* (Kuhn, 2008). For each pair of algorithms, the analysis was repeated twice such that each variable of the pair served once as the reference and once as the predictor. Following measures were used to evaluate pairwise concordance between different algorithms for probable PDD: 1) Cohen's κ with its 95% confidence interval (CI) computed via the `cohen.kappa()` function from the R package *psych* (William Revelle, 2024); 2) Accuracy (i.e., the proportion of correct predictions, both true positives and true negatives, among the total number of cases) with its 95% CI; 3) Sensitivity/Recall (i.e., the proportion of true positives); and 4) Specificity (i.e., the proportion of true negatives).

Finally, the No Information Rate (NIR) was calculated for each pair of algorithms. NIR is the accuracy that could be obtained by always predicting the majority class and in our case it is equivalent to the complement of the PDD rate estimate according to the reference algorithm. Accuracy of prediction was compared to the NIR via a one-sided Exact Binomial Test as implemented by the `binom.test()` R stats function. Reference/predictor pairs associated with p < .05 were considered to show significantly better accuracy than NIR. In other words, for reference/predictor pairs associated with p < .05, we conclude that knowing the probable PDD status according to the predictor algorithm helps to estimate the probable PDD status according to the reference algorithm and the two algorithms thus show substantial concordance.

Missing data were handled by complete cases analysis and pairwise complete cases analysis. In other words, each univariate and analysis pairwise comparison used all available data. Data wrangling and visualizations were done in the *tidyverse* package (Wickham et al., 2019) and tables were formatted in the *gt* package (Iannone et al., 2024). All analyses were conducted within the R (version 4.3.3) software environment for statistical computing (R Core Team, 2024). The software code supporting this article is available at https://github.com/josefmana/demcrit.git.

Results

Sample Description

Total of 204 patients **were considered for the study, out of which one patient was excluded due to missing neuropsychological data, resulting in a final sample of 203 patients**.

The sample included 126 (62%) men, with an average of 59.01 (SD = 8.37) years of age, 13.77 (SD = 3.08) years of education, 10.78 (SD = 4.25) years of disease duration, 37.42 (SD = 12.99) Unified Parkinson Disease Rating Scale (UPDRS), part III in medication OFF state and 15.83 (SD = 8.08) Unified Parkinson Disease Rating Scale (UPDRS), part III in medication ON state. **Descriptive statistics for neuropsychiatric symptoms indicated within average level of depressive and anxiety symptoms in our cohort, with the average BDI-II of 10.79 (SD = 7.02), average STAI X1 of 38.94 (SD = 8.97), and average STAI X2 of 40.34 (SD = 7.78). However, according to the psychiatric assessment, none of the patients with probable PDD was suffering from the major depressive disorder, delirium or other neuropsychiatric abnormalities that would excluded the diagnosis.** Cognitive characteristics of the sample are summarized in Table 2.

[Insert Table 2 here]

PDD Rate Estimates

Algorithm-wise rate of PDD estimates **corresponding numbers of patients with available data** are presented in Table A3. **In total, there were 2 algorithms based on 194 complete observations (4.4% missing). One algorithm used 199, 200, and 201 complete observations, respectively (2%, 1.5% and 1% missing). A further 32 algorithms were based on 202 complete observations (0.5% missing), and 31 algorithms included the full sample of 203 patients.**

On average, the estimated PDD rate was 6.03% (SD = 3.45, Md = 3.96, range 2.01-16.75). Notably, the estimates were substantially lower when FAQ item 9 was used as a criterion of IADL deficit (M = 3.12% SD = 0.48, Md = 2.97, range 2.01-3.96) compared to using the total FAQ score criterion (M = 8.94% SD = 2.54, Md = 9.11, range 3.50-16.75) as demonstrated in Figure 1 **(see also Figure A1 for per-operationalisation distribution of PDD rate estimates**). Neither age, sex nor their interaction (*p*s ≥ .112) reliably predicted probable PDD classification across algorithms (see Figure A2 and Figure A3).

[Insert Figure 1 here]

Concordance between Algorithms

Results of the analyses of prediction Accuracy, Cohen's κ, Sensitivity and Specificity are presented in Figure 2, Figure A4, Figure A5 and Figure A6 respectively. **The number of complete pairwise cases ranged from 191 (5.91% missing) to 203 (full sample). Most comparisons (89.90%) were based off of 200 or more observations.**[1]

Generally, algorithms that employed the same operationalization of IADL deficit showed substantial pairwise concordance, however, algorithms that operationalized IADL deficit differently did not. Whereas among algorithms with identical IADL deficit operationalization, the agreement judged by Cohen's κ was moderately high (operationalization by FAQ total score: κ = 0.75, SD = 0.13; operationalization by FAQ item 9: κ = 0.86, SD = 0.09), among algorithms that differ in IADL deficit operationalization but are otherwise identical it was low: κ = 0.43, SD = 0.08.

[Insert Figure 3 here]

Prediction of Level II Criteria

For easier interpretability of our results, we next examined cases where Level II algorithms served as a reference and Level I algorithms as a predictor. Table 3 shows five Level I algorithms with the highest and five with the lowest accuracy in predicting Level II classification of probable PDD.

When IADL deficit was defined by total FAQ score, the Level II estimate of PDD rate was 10.31%. All five Level I algorithms that approximated the Level II classification most accurately were MoCA-based and defined Executive Function deficit by Clock drawing rather than Verbal fluency test. On the other hand, two out of the five Level I algorithms with the lowest accuracy

[1] **Due to the large number of entries (4624 rows x 21 columns representing pairwise comparisons and metrics of interest respectively), the table with numerical results is not presented here or in the Appendix. Instead, we share the table share as data in the accompanying R package available at https://github.com/josefmana/demcrit.git. To obtain the table in format not dependent on R, follow the tutorial at https://josefmana.github.io/demcrit/articles/concordance.html**

were MMSE-based, whereas the remaining three were MoCA-based and defined Executive function deficit by Verbal fluency test.

When IADL deficit was defined by FAQ item 9 score, the Level II estimate of PDD rate was 3.61%. Overall, the difference between the most accurate and the least accurate Level I algorithms was lower than in the case of IADL deficit being defined by FAQ total score (see Table 3). The five most accurate algorithms were all MoCA-based, defined Executive Function deficit by Clock drawing (with threshold < 2) and in majority of cases defined Language deficit by Animal naming. Two out of the five Level I algorithms with the lowest accuracy were MMSE-based, whereas the remaining three were MoCA-based and defined Executive Function deficit by Clock drawing (with threshold < 3) and Language deficit by Abstraction.

Finally, suppose the predictors are sorted by their balanced accuracy (i.e., average of sensitivity and specificity) instead of raw accuracy. In that case, the results are similar, with the exception that for the prediction of Level II with total FAQ score algorithm for probable PDD, the highest balanced accuracy was achieved by the sMoCA algorithm with sensitivity 0.95 and specificity 0.93 (see Table A4).

Discussion

This study systematically investigated the application of multiple Level I PDD diagnostic criteria. Our results show variability in PDD rate estimates, strongly influenced by the choice of cognitive screening instrument (MMSE, MoCA and sMoCA) and the operationalization of functional impairment. The divergence observed across algorithms demonstrates the sensitivity of diagnostic outcomes to seemingly negligible methodological choices.

Variability in PDD Rate Estimates

Our results showed a wide range in estimated PDD rate across algorithms, ranging from 2.01% to 16.75%. Estimates reached lower rates when using solely FAQ item 9 (as an approximation of the pill questionnaire suggested by Dubois et al. (2007)) in comparison with the full FAQ scale. This discrepancy highlights the diagnostic importance of how IADLs are assessed.

Our overall PDD rates were consistently lower than previous studies regarding PDD among PD patients, demonstrating wide variability based on various criteria used. For instance, a retrospective study reported a PDD rate of 19.7% (Rana et al., 2011), while other clinical investigation found even higher rate, reaching up to 30% (Aarsland et al., 2005). A recent meta-analysis synthesizing global data placed the expected PDD rate in PD at 26.30% (Sousa et al., 2022). Compared to these estimates, our study reports generally lower PDD rates, likely reflecting differences in diagnostic criteria, methodology and sample characteristics. Specifically, our sample was younger compared to other PD cohorts and age was repeatedly shown to be a strong predictor of PDD across studies (Rana et al., 2011; Sousa et al., 2022).

Interestingly, we did not observe any reliable age-related differences in PDD rate within our cohort. This lack of age-dependency may, however, be also stemming from the relatively younger age of our cohort, because previous reports indicate that association between age and PDD is not linear but increases with age and may not reach substantial values before **older age. In both Rana et al. (2011) and Oh et al. (2016), nine out of ten patients with dementia were 70 years of age or older. In our sample, only 4.5% participants were in this age range. Consequently, studies with older cohorts are probably necessary to detect a robust association between age and the risk of probable PDD.**

Concordance Between Diagnostic Algorithms

Pairwise comparisons of diagnostic algorithms showed that agreement was notably stronger among those using the same IADL operationalization compared to those using different IADL definitions. Moreover, the agreement was slightly higher between algorithms that defined IADL deficit by FAQ item 9 compared to algorithms that defined it using the full FAQ scale. One possible explanation of this difference follows from the observation that algorithms using the full-scale definition yielded higher PDD rate estimates. Because there was a higher probability of being diagnosed with IADL deficit based on the full FAQ scale, there was also a bigger room for disagreement in the Cognitive Impairment status when different indexes were used (e.g. by defining executive deficit via clock drawing vs. verbal fluency).

Overall, when the same IADL definitions were used across algorithms, we observed concordance levels varying from moderate (using FAQ total score) to strong (using FAQ item 9), consistent with inter-rater reliability analysis (McHugh, 2012). Contrarily, the concordance between algorithms using different IADL deficit definitions was equivalent to minimal agreement. This demonstrates that even slight methodological differences can yield divergent diagnostic outcomes. Such findings are critical for clinicians relying on Level I criteria for eligibility decisions, as the choice of algorithm could lead to contradictory classifications of PDD status.

Predictive Validity Comparison With Level II Criteria

Using Level II diagnosis as the gold standard, MoCA-based Level I algorithms, particularly those using Clock Drawing to assess executive function, demonstrated the highest predictive accuracy. This supports recent proposals to modernise PDD diagnostic frameworks (Kulisevsky et al., 2024), favouring MoCA-derived components and more specific, PD-tailored functional assessment tools. In contrast, MMSE-based algorithms consistently underperformed, suggesting limited sensitivity in capturing cognitive deficits typical in PDD.

Furthermore, in the algorithm using sMoCA, the raw accuracy was moderate, however, the balanced accuracy (i.e. combined sensitivity and specificity) was high. **Consequently, sMoCA appears particularly suitable for approximating Level II PDD diagnosis in populations that differ in PDD prevalence from our cohort, since balanced accuracy, unlike raw accuracy, is independent of prevalence in the sample. Moreover, because the sMoCA algorithm demonstrated higher sensitivity while maintaining comparable specificity (see Table A4), it may be especially valuable in contexts where false negatives carry substantial clinical cost. In such cases, a neuropsychologist might use sMoCA as an initial screening tool and proceed to a full Level II assessment only for patients who meet criteria for probable PDD in this preliminary stage.**

Constraints on Generality

This study’s generalizability is limited by the homogeneity of the patient cohort, which does not reflect the diversity of cognitive profiles seen in broader PD populations. Specifically, the younger age of the sample, and a possible underrepresentation of high-risk

phenotypes for PDD constraints the generality of the presented findings.

As noted above, the younger age of our sample may partly explain the lower rate of PDD observed compared to previous studies. As discussed in the *Theoretical and Empirical Estimands* section of the Appendix, neither estimates of PDD rate nor predictive performance of demographic variables should therefore not be generalized beyond PD patients that are candidates for DBS. The extent to which our findings on the concordance between diagnostic algorithms apply to the broader PD population remains to be determined in future studies using different types of cohorts, such as de novo patients or community-based samples.

Longitudinal cohort studies suggest that patients considered for DBS represent a subset of the PD population with a distinct cognitive phenotype (Bove et al., 2020; Mana et al., 2024). Specifically, findings of gradual post-surgical cognitive decline predicted by pre-surgical executive deficits indicate that DBS candidates may be preferentially drawn from a fronto-striatal phenotype characterized by slowly progressing executive dysfunction, rather than from a posterior phenotype marked by visuospatial impairment (Kehagia et al., 2012). Importantly, patients with the posterior phenotype may be at greater risk of developing PDD within as little as five years after disease onset (Summers et al., 2024; Williams-Gray et al., 2009).

In our study, visuospatial function was assessed uniformly across all algorithms within a given screening measure (MoCA cube or MMSE pentagons). By contrast, we compared two operationalizations of executive dysfunction, the clock drawing test and verbal fluency. The clock drawing test showed stronger predictive value for level II diagnosis than verbal fluency, suggesting that even within our cohort, patients who developed PDD may exhibit features of the posterior phenotype.

However, the use of a DBS cohort also offered several methodological advantages. All patients underwent standardized and comprehensive neuropsychological testing, resulting in a well-characterized dataset that enabled a systematic evaluation of multiple diagnostic

algorithms. Moreover, because dementia is a common exclusion criterion for DBS treatment (Bronstein et al., 2011), examining the diagnostic accuracy of algorithms for PDD in a pre-DBS cohort is informative in its own right.

Limitations and Future Directions

Due to the retrospective nature of the study, some patients lacked one or more key measures required for the diagnosis of probable PDD by certain algorithms. Missing data were handled using the pairwise complete cases method. The main advantages of this approach are its straightforward implementation and preservation of statistical power. However, it may introduce bias, particularly when data are not missing completely at random (MCAR) and when causal inference is the goal (Little & Rubin, 2019).

Modern alternatives, such as multiple imputation techniques, have been shown to produce less biased estimates than case deletion strategies in analyses based on confusion matrices (Karakaya et al., 2014), whereas the evidence supporting their superiority in the estimation of Cohen's κ remains limited (De Raadt et al., 2019). These modern approaches also require careful specification of the imputation model and the underlying causal mechanism of missingness to ensure appropriate covariate selection (Bianco et al., 2023; Long et al., 2011). For the sake of parsimony, we did not employ advanced missing-data techniques but instead explicitly described the observed missingness patterns. Consequently, our findings should be regarded as exploratory and primarily serve as a basis for hypothesis generation in future, more controlled studies.

Another possible avenue for analysing the present data would be to exploit their hierarchical structure by fitting a multilevel model with algorithms nested within cognitive tests (see Figure 1). Such an approach could leverage partial pooling (Gelman et al., 2012) and might approximate modern item response theory frameworks (Bürkner, 2021), where PDD is conceptualised as a latent trait and the individual diagnostic algorithms act as items measuring it. If feasible, this psychometric perspective may represent an exciting new direction for conceptualising complex clinical phenomena such as PDD (cf. Miller et al.

(2012), Kiselica and Benge (2021)).

An additional limitation concerns the use of the FAQ questionnaire for IADL assessment. The FAQ is a subjective or informant-reliant measure and thus susceptible to bias. Moreover, its content can vary across sociocultural contexts, limiting cross-cultural transferability. For example, activities such as financial management, cooking or driving, are not universally practised across societies. Consequently, both the FAQ scores and diagnostic thresholds used to indicate IADL impairment may not be directly transferable between cultural settings (O'Donald & Calia, 2025). These factors may influence both the sensitivity and ecological validity of the functional criteria used (Bezdicek et al., 2011; Bezdicek, Stepankova, et al., 2016).

Furthermore, IADL measures may correlate with neuropsychological results of specific cognitive domains, particularly attention/processing speed, and executive function (Moheb et al., 2017; Reppermund et al., 2011). If such correlations were due to shared error variance, they could bias concordance indeces measured in this study. To examine this possibility, we conducted a post-hoc simulation experiment, available at https://josefmana.github.io/demcrit/articles/correlation.html. The results indicated that correlations between IADL and neuropsychological measures may either increase or decrease accuracy and balanced accuracy, while consistently inflating Cohen's κ estimates. However, given the correlations observed in our dataset, the effect size was small and unlikely to alter the conclusions of our study.

To address concerns about measuring IADL outlined above, future research should consider using PD-specific questionnaires or more objective tools. Promising options include the Penn Parkinson's Daily Activities Questionnaire-15 (Brennan et al., 2016), questionnaire adaptations including items regarding gadgets use and digital literacy (Postema et al., 2024) or performance-based assesssments (Schmitter-Edgecombe et al., 2020). Our findings underscore the importance of IADL measurement for the PDD diagnosis. Therefore, we recommend exploring more reliable tools with high ecological validity.

Finally, whereas our study systematically investigated how varying definition of global deficit, impaired cognition and IADL deficit affect probable PDD classification, it did not explore associations of PDD diagnosis with its neuropsychiatric (e.g. anxiety profile) and biomarker correlates. **Instead, we only ensured the absence of acute or severe psychiatric symptoms that would preclude a diagnosis of probable PDD, as verified through assessment by a trained neuropsychiatrist. However, both the current diagnostic criteria for PDD (Dubois et al., 2007) and recent proposals for their revision (Kulisevsky et al., 2024) emphasize the use of standardized psychometric instruments for assessing neuropsychiatric symptoms. Incorporating such measures could enhance both the efficiency and transparency of the diagnostic process. Future research should therefore investigate how integrating structured neuropsychiatric assessments and biomarker data may refine PDD diagnostic accuracy and improve clinical utility.**

Conclusions

In sum, our study systematically investigated how varying definitions of impaired cognition and IADL deficit affect PDD diagnostic accuracy. Our study highlights the variability in PDD classification across Level I diagnostic algorithms, significantly influenced by IADL operationalisation and the choice of cognitive screening tools. The findings strongly support the call for a change of the current diagnostic criteria (Kulisevsky et al., 2024), favouring the use of MoCA-based components and comprehensive IADL assessments.

Conservative criteria, such as reliance on pill questionnaire (i.e. FAQ item 9 equivalence), may fail to detect functional decline and thus under-identify true cases of PDD. Importantly, concordance across algorithms rises significantly, reaching moderate to high values, when the same definition of IADL is used (either FAQ total or FAQ item 9). Moreover, when using MoCA-based algorithms instead of MMSE-based ones, we can observe better approximations to the Level II battery.

Future studies should aim to replicate our results on larger and different cohorts **with more heterogenous sample, and to explore how varying operationalization of other**

components of probable PDD, namely psychiatric symptoms, affects PDD rates and concordance between diagnostic algorithms. To make this process easier, the code used to generate our results is publicly available and easily applicable to similarly structured data.

References

Aarsland, D., Zaccai, J., & Brayne, C. (2005). A systematic review of prevalence studies of dementia in parkinson's disease. *Movement Disorders: Official Journal of the Movement Disorder Society*, *20*(10), 1255–1263. https://doi.org/10.1016/j.parkreldis.2011.05.017

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

Benedict, R. H. B. (1997). *Brief visuospatial memory test revised: Professional manual*. Psychological Assessment Resources.

Benton, A. L., Varney, N. R., & Hamsher, K. (1983). Visuospatial judgment: A clinical test. *Archives of Neurology*, *40*(3), 429–432. https://doi.org/10.1001/archneur.1978.00500300038006

Bezdicek, O., Červenková, M., Moore, T. M., Georgi, H. S., Sulc, Z., Wolk, D. A., Weintraub, D. A., Moberg, P. J., Jech, R., Kopecek, M., & Roalf, D. R. (2020). Determining a short form Montreal Cognitive Assessment (s-MoCA) Czech version: Validity in mild cognitive impairment Parkinson's disease and cross-cultural comparison. *Assessment*, *27*(8), 1960–1970. https://doi.org/10.1177/1073191118778896

Bezdicek, O., Georgi, H., Nikolai, T., & Kopeček, M. (2021). *Pražská verze stroopova testu*. Karolinum. https://karolinum.cz/en/books/bezdicek-prazska-verze-stroopova-testu-25158

Bezdicek, O., Lukavský, J., & Preiss, M. (2011). Functional activities questionnaire, czech version – a validation study. *Česká a Slovenská Neurologie a Neurochirurgie*, *74*(107), 36–42. https://www.csnn.eu/casopisy/ceska-slovenska-neurologie/2011-1/validizacni-studie-ceske-verze-dotazniku-faq-34140

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., Nikolai, T., Michalec, J., Růžička, F., Havránková, P., Roth, J., Jech, R., & Růžička, E. (2016). The Diagnostic Accuracy of Parkinson's Disease Mild Cognitive Impairment Battery Using the Movement Disorder Society Task Force Criteria. *Movement Disorders Clinical Practice*, *4*(2), 237–244. https://doi.org/10.1002/mdc3.12391

Bezdicek, O., Stepankova, H., Martinec Novakova, L., & Kopecek, M. (2016). Toward the processing speed theory of activities of daily living in healthy aging: Normative data of the functional activities questionnaire. *Aging Clinical and Experimental Research*, *28*, 239–247.

Bezdicek, O., Stepankova, H., Moták, L., Axelrod, B. N., Woodard, J. L., Preiss, M., Nikolai, T., Růžička, E., & Poreh, A. (2013). 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., & Růžička, 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

Bianco, A. M., Boente, G., GonzálezManteiga, W., & PérezGonzález, A. (2023). Estimators for ROC curves with missing biomarkers values and informative covariates. *Statistical Methods & Applications*, *32*(3), 931–956. https://doi.org/10.1007/s10260-022-00680-z

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), e384–e392. https://doi.org/10.1212/WNL.0000000000009822

Brennan, L., Siderowf, A., Rubright, J. D., Rick, J., Dahodwala, N., Duda, J. E., Hurtig, H., Stern, M., Xie, S. X., Rennert, L., Karlawish, J., Shea, J. A., Trojanowski, J. Q., & Weintraub, D.

(2016). The Penn Parkinson's Daily Activities Questionnaire-15: Psychometric properties of a brief assessment of cognitive instrumental activities of daily living in Parkinson's disease. *Parkinsonism & Related Disorders*, *25*, 21–26. https://doi.org/10.1016/j.parkreldis.2016.02.020

Bronstein, J. M., Tagliati, M., Alterman, R. L., Lozano, A. M., Volkmann, J., Stefani, A., Horak, F. B., Okun, M. S., Foote, K. D., Krack, P., Pahwa, R., Henderson, J. M., Hariz, M. I., Bakay, R. A., Rezai, A., Marks, W. J., Moro, E., Vitek, J. L., Weaver, F. M., … DeLong, M. R. (2011). Deep Brain Stimulation for Parkinson Disease. *Archives of Neurology*, *68*(2). https://doi.org/10.1001/archneurol.2010.260

Bürkner, P.-C. (2021). Bayesian item response modeling in r with brms and stan. *Journal of Statistical Software*, *100*(5), 1–54. https://doi.org/10.18637/jss.v100.i05

Cereda, E., Cilia, R., Klersy, C., Siri, C., Pozzi, B., Reali, E., Colombo, A., Zecchinelli, A. L., Mariani, C. B., Tesei, S., Canesi, M., Sacilotto, G., Meucci, N., Zini, M., Isaias, I. U., Barichella, M., Cassani, E., Goldwurm, S., & Pezzoli, G. (2016). Dementia in parkinson's disease: Is male gender a risk factor? *Parkinsonism & Related Disorders*, *26*, 67–72. https://doi.org/10.1016/j.parkreldis.2016.02.024

Ciharova, M., Cígler, H., Dostálová, V., Šivicová, 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

De Raadt, A., Warrens, M. J., Bosker, R. J., & Kiers, H. A. L. (2019). Kappa Coefficients for Missing Data. *Educational and Psychological Measurement*, *79*(3), 558–576. https://doi.org/10.1177/0013164418823249

Dubois, B., Burn, D., Goetz, C., Aarsland, D., Brown, R. G., Broe, G. A., Dickson, D., Duyckaerts, C., Cummings, J., Gauthier, S., Korczyn, A., Lees, A., Levy, R., Litvan, I., Mizuno, Y., McKeith, I. G., Olanow, C. W., Poewe, W., Sampaio, C., … Emre, M. (2007). Diagnostic procedures for Parkinson's disease dementia: Recommendations from the movement disorder society task force. *Movement Disorders*, *22*(16), 2314–2324. https://doi.org/10.1002/mds.21844

Emre, M., Aarsland, D., Brown, R., Burn, D. J., Duyckaerts, C., Mizuno, Y., Broe, G. A., Cummings, J., Dickson, D. W., Gauthier, S., Goldman, J., Goetz, C., Korczyn, A., Lees, A., Levy, R., Litvan, I., McKeith, I., Olanow, W., Poewe, W., … Dubois, B. (2007). Clinical diagnostic criteria for dementia associated with Parkinson’s disease. *Movement Disorders*, *22*(12), 1689–1707. https://doi.org/10.1002/mds.21507

Folstein, M. F., Folstein, S. E., & McHugh, P. R. (1975). “Mini-mental state”. *Journal of Psychiatric Research*, *12*(3), 189–198. https://doi.org/10.1016/0022-3956(75)90026-6

Frydrychová, Z., Kopeček, M., Bezdicek, O., & Štěpánková, J. (2018). České normy pro revidovaný reyův auditorně-verbální test učení (RAVLT) pro populaci starších osob. *Československá Psychologie*, *62*(4), 330–349. https://cejsh.icm.edu.pl/cejsh/element/bwmeta1.element.02da6923-6394-4d8d-b018-dd6e3b2503be

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

Havlík, F., Mana, J., Dušek, P., Jech, R., Ržička, E., Kopeček, M., Georgi, H., & Bezdicek, O. (2020). Brief visuospatial memory test-revised: Normative data and clinical utility of learning indices in parkinson’s disease. *Journal of Clinical and Experimental Neuropsychology*, *42*(10), 1099–1110.

Iannone, R., Cheng, J., Schloerke, B., Hughes, E., Lauer, A., Seo, J., Brevoort, K., & Roy, O. (2024). *Gt: Easily create presentation-ready display tables*. https://CRAN.R-project.org/package=gt

Kaplan, E., Goodglass, H., & Weintraub, S. (1983). *Boston naming test*. Lea & Febiger.

Karakaya, J., Karabulut, E., & Yucel, R. M. (2014). Sensitivity to imputation models and assumptions in receiver operating characteristic analysis with incomplete data. *Journal of Statistical Computation and Simulation*, *85*(17), 3498–3511.

https://doi.org/10.1080/00949655.2014.983111

Kehagia, A. A., Barker, R. A., & Robbins, T. W. (2012). Cognitive Impairment in Parkinson's Disease: The Dual Syndrome Hypothesis. *Neurodegenerative Diseases*, *11*(2), 79–92. https://doi.org/10.1159/000341998

Kiselica, A. M., & Benge, J. (2021). An item response theory approach to clinical diagnosis of MCI and dementia: Illustrations from ADNI and NACC data. *Alzheimer's & Dementia*, *17*(S6), e049773. https://doi.org/https://doi.org/10.1002/alz.049773

Kopecek, M., Stepankova, H., Lukavsky, J., Ripova, D., Nikolai, T., & Bezdicek, O. (2016). Montreal cognitive assessment (MoCA): Normative data for old and very old Czech adults. *Applied Neuropsychology: Adult*, *24*(1), 23–29. https://doi.org/10.1080/23279095.2015.1065261

Kuhn, M. (2008). Building predictive models in r using the caret package. *Journal of Statistical Software*, *28*(5), 1–26. https://doi.org/10.18637/jss.v028.i05

Kulisevsky, J., Litvan, I., Weintraub, D., Goldman, J. G., Tröster, A. I., Lewis, S. J., Parkinson, I., Group, M. D. S. P.-M. S., et al. (2024). A call for change: Updating the operational definition for dementia in parkinson's disease. *Movement Disorders Clinical Practice*, *12*(3), 296–301. https://doi.org/10.1002/mdc3.14305

Little, R., & Rubin, D. (2019). Complete-case and available-case analysis, including weighting methods. In *Statistical analysis with missing data, third edition* (pp. 47–66). John Wiley & Sons, Ltd. https://doi.org/https://doi.org/10.1002/9781119482260.ch3

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

Long, Q., Zhang, X., & Hsu, C.-H. (2011). Nonparametric multiple imputation for receiver operating characteristics analysis when some biomarker values are missing at random.

Statistics in Medicine, *30*(26), 3149–3161. https://doi.org/10.1002/sim.4338

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

Mana, J., Bezdicek, O., Růžička, F., Lasica, A., Šmídová, A., Klempířová, O., Nikolai, T., Uhrová, T., Růžička, E., Urgošík, D., & Jech, R. (2024). Preoperative cognitive profile predictive of cognitive decline after subthalamic deep brain stimulation in parkinson's disease. *European Journal of Neuroscience*, *60*(7), 5764–5784. https://doi.org/https://doi.org/10.1111/ejn.16521

McHugh, M. L. (2012). Interrater reliability: The kappa statistic. *Biochemia Medica*, 276–282. https://doi.org/10.11613/bm.2012.031

Meireles, J., & Massano, J. (2012). Cognitive impairment and dementia in parkinsons disease: Clinical features, diagnosis, and management. *Frontiers in Neurology*, *3*. https://doi.org/10.3389/fneur.2012.00088

Miller, T. M., Balsis, S., Lowe, D. A., Benge, J. F., & Doody, R. S. (2012). Item response theory reveals variability of functional impairment within clinical dementia rating scale stages. *Dementia and Geriatric Cognitive Disorders*, *32*(5), 362–366. https://doi.org/10.1159/000335727

Moheb, N., Mendez, M. F., Kremen, S. A., & Teng, E. (2017). Executive Dysfunction and Behavioral Symptoms Are Associated with Deficits in Instrumental Activities of Daily Living in Frontotemporal Dementia. *Dementia and Geriatric Cognitive Disorders*, *43*(1-2), 89–99. https://doi.org/10.1159/000455119

Mullner, J., Ruisl, I., & Farkas, G. (1980). *Dotaznik na meranie uzkosti a uzkostlivosti - STAI*. Bratislava: Psychodiagnostické a didaktické testy.

Nasreddine, Z. S., Phillips, N. A., Bédirian, V., Charbonneau, S., Whitehead, V., Collin, I., Cummings, J. L., & Chertkow, H. (2005). The Montreal Cognitive Assessment, MoCA: A Brief Screening Tool For Mild Cognitive Impairment. *Journal of the American Geriatrics Society*, *53*(4), 695–699. https://doi.org/10.1111/j.1532-5415.2005.53221.x

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

O’Donald, F., & Calia, C. (2025). The Process of Translation and Cross-Cultural Adaptation of Functional Assessment Tools for Dementia: A Systematized Review. *Health Science Reports*, *8*(1). https://doi.org/10.1002/hsr2.70289

Oh, Y.-S., Kim, J.-S., Park, I.-S., Shim, Y.-S., Song, I.-U., Park, J.-W., Lee, P.-H., Lyoo, C.-H., Ahn, T.-B., Ma, H.-I., et al. (2016). Prevalence and treatment pattern of parkinson’s disease dementia in korea. *Geriatrics & Gerontology International*, *16*(2), 230–236.

Pfeffer, R. I., Kurosaki, T. T., Harrah, C. H., Chance, J. M., & Filos, S. (1982). Measurement of Functional Activities in Older Adults in the Community. *Journal of Gerontology*, *37*(3), 323–329. https://doi.org/10.1093/geronj/37.3.323

Postema, M. C., Dubbelman, M. A., Claesen, J., Ritchie, C., Verrijp, M., Visser, L., Visser, P.-J., Zwan, M. D., Flier, W. M. van der, & Sikkes, S. A. M. (2024). Facilitating clinical use of the Amsterdam Instrumental Activities of Daily Living Questionnaire: Normative data and a diagnostic cutoff value. *Journal of the International Neuropsychological Society*, *30*(6), 615–620. https://doi.org/10.1017/s1355617724000031

Postuma, R. B., Berg, D., Stern, M., Poewe, W., Olanow, C. W., Oertel, W., Obeso, J., Marek, K., Litvan, I., Lang, A. E., Halliday, G., Goetz, C. G., Gasser, T., Dubois, B., Chan, P., Bloem, B. R., Adler, C. H., & Deuschl, G. (2015). MDS clinical diagnostic criteria for Parkinson’s disease. *Movement Disorders*, *30*(12), 1591–1601. https://doi.org/10.1002/mds.26424

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

Rana, A. Q., Yousuf, M. S., Naz, S., & Qa’aty, N. (2011). Prevalence and relation of dementia to various factors in parkinson’s disease. *Psychiatry and Clinical Neurosciences*, *65*(4), 317–321. https://doi.org/10.1111/j.1440-1819.2011.02291.x

Reitan, R. (2004). The Trail Making Test as an initial screening procedure for neuropsychological impairment in older children. *Archives of Clinical Neuropsychology*, *19*(2), 281–288. https://doi.org/10.1016/s0887-6177(03)00042-8

Reppermund, S., Sachdev, P. S., Crawford, J., Kochan, N. A., Slavin, M. J., Kang, K., Trollor, J. N., Draper, B., & Brodaty, H. (2011). The relationship of neuropsychological function to instrumental activities of daily living in mild cognitive impairment. *International Journal of Geriatric Psychiatry*, *26*(8), 843–852. https://doi.org/https://doi.org/10.1002/gps.2612

Rey, A. (1964). *L'examen clinique en psychologie (the clinical psychological examination)*. Presses Universitaires de France.

Roalf, D. R., Moore, T. M., Mechanic-Hamilton, D., Wolk, D. A., Arnold, S. E., Weintraub, D. A., & Moberg, P. J. (2017). Bridging cognitive screening tests in neurologic disorders: A crosswalk between the short Montreal Cognitive Assessment and Mini-Mental State Examination. *Alzheimer's & Dementia*, *13*(8), 947–952. https://doi.org/10.1016/j.jalz.2017.01.015

Royall, D. R., Cordes, J. A., & Polk, M. (1998). CLOX: an executive clock drawing task. *Journal of Neurology, Neurosurgery & Psychiatry*, *64*(5), 588–594. https://doi.org/10.1136/jnnp.64.5.588

Schmitter-Edgecombe, M., Cunningham, R., McAlister, C., Arrotta, K., & Weakley, A. (2020). The night out task and scoring application: an ill-structured, open-ended clinic-based test representing cognitive capacities used in everyday situations. *Archives of Clinical Neuropsychology*, *36*(4), 537–553. https://doi.org/10.1093/arclin/acaa080

Sousa, C. S. e, Alarcão, J., Martins, I. P., & Ferreira, J. J. (2022). Frequency of dementia in parkinson's disease: A systematic review and meta-analysis. *Journal of the Neurological Sciences*, *432*, 120077. https://doi.org/10.1016/j.jns.2021.120077

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.

Stepankova, H., Nikolai, T., Lukavsky, J., Bezdicek, O., Vrajova, M., & Kopecek, M. (2015).

Mini-Mental State Examination – česká normativní studie. *Ceska a slovenska neurologie a neurochirurgie*, *78*(111), 57–63.

Summers, D., Spencer, K., Okasaki, C., & Huber, J. E. (2024). An examination of cognitive heterogeneity in parkinson disease: The dual-syndrome hypothesis. *Journal of Speech, Language, and Hearing Research*, *67*(4), 1127–1135. https://doi.org/10.1044/2024/_JSLHR-23-00621

Wechsler, D. (1997). *Wechsler adult intelligence scale—third edition (WAIS-III)*. Psychological Corporation. https://books.google.cz/books/about/Wais_III_Wechsler_Adult_Intelligence_Sca.html?id=qTCuGQAACAAJ

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., … Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, *4*(43), 1686. https://doi.org/10.21105/joss.01686

William Revelle. (2024). *Psych: Procedures for psychological, psychometric, and personality research*. Northwestern University. https://CRAN.R-project.org/package=psych

Williams-Gray, C. H., Evans, J. R., Goris, A., Foltynie, T., Ban, M., Robbins, T. W., Brayne, C., Kolachana, B. S., Weinberger, D. R., Sawcer, S. J., & Barker, R. A. (2009). The distinct cognitive syndromes of Parkinson's disease: 5 year follow-up of the CamPaIGN cohort. *Brain*, *132*(11), 2958–2969. https://doi.org/10.1093/brain/awp245

Zemanová, N., Bezdíček, O., Michalec, J., Nikolai, T., Roth, J., Jech, R., & Ržička, E. (2016). Validační studie české verze bostonského testu pojmenování. *Česká a Slovenská Neurologie a Neurochirurgie*, *79*(112), 3. https://doi.org/10.14735/amcsnn2016307

Table 1

Summary of probable PDD operationalizations compared in the study.

Type	Global functioning	Attention	Executive Function
MMSE-based	MMSE < 26	Sevens backwards < 4	Clock drawing < 2 OR Lexical fluency (S) < 10
MoCA-based	MoCA < 27	Sevens backwards < 3	Clock drawing < {2, 3} OR Lexical fluency (K) <
sMoca-based	sMoCA < 13	-	-
Level II	-	TMT A & WAIS DSB	CF A & PST C

[a] The visual memory was evaluated based on WMS-III Family Pictures or BVMTR depending on which test was used in the assessment. This lead to no missing values because each patient underwent assessment via one of these tests.

Note. MMSE: Mini-Mental State Examination; MoCA: Montreal Cognitive Assessment; sMoCA: short version of the MoCA; TMT A: Trail Making Test, Part A; WAIS DSB: Wechsler Adult Intelligence Scale Digit Span, Backwards; CF A: Categorical Verbal Fluency, Animals; PST C: Prague Stroop Test, Colours; WAIS Similarities: Wechsler Adult Intelligence Scale, Similarities; BNT 60: Boston Naming Test; RAVLT delayed recall: Rey Auditory Verbal Learning Test, Delayed Recall; BVMTR delayed recall: Brief Verbal Memory Test, Delayed Recall; WMS-III Family Pictures: Wechsler Memory Scale Family Pictures; JoLO: Boston Judgement of Line Orientation; CLOX: Clock Drawing Test. The OR operator implies that exactly one of the criteria listed is utilized within a single operationalization; the & operator implies that both criteria are used at the same time within a single operationalization; each threshold value within the set brackets { } was used to define probable PDD once in combination with all the other criteria on the same row.

Table 2

Cognitive characteristics of the sample.

	N	Md	Min-max	M	SD
MMSE					
Total score (Range 0-30)	203	27	15-30	26.69	2.22
Sevens (Range 0-5)	1/2/8/20/34/139	-	-	-	-
VF S (Number of Words per Minute)	202	15	1-34	14.95	5.80
Clock Drawing (Range 0-2)	26/91/86	-	-	-	-
Pentagons (Range 0-1)	187 (92%)	-	-	-	-
Three words (Range 0-3)	5/14/60/124	-	-	-	-
MoCA					
Total score (Range 0-30)	203	24	9-30	24.07	3.48
sMoCA total score (Range 0-16)	203	11	1-16	11.26	2.74
Sevens (Range 0-3)	1/2/29/171	-	-	-	-
VF K (Number of Words per Minute)	204	16	0-29	15.49	5.34
Clock drawing (Range 0-3)	24/83/96	-	-	-	-
Cube drawing (Range 0-1)	164 (81%)	-	-	-	-
Five words (Range 0-5)	69/19/29/39/22/25	-	-	-	-
Animal naming (Range 0-3)	10/193	-	-	-	-
Abstraction (Range 0-2)	7/72/124	-	-	-	-
Affect					
BDI (Range 0-63)	203	10	0-34	10.79	7.02
STAI X1 (Range 0-80)	186	38	20-72	38.94	8.97
STAI X2 (Range 0-80)	184	40	22-63	40.34	7.78
IADL					
FAQ (Range 0-30)	203	2	0-25	4.05	4.89
FAQ 9 (Range 0-1)	144/47/10/1	-	-	-	-

Table 3

Level I algorithms for probable PDD as predictors of Level II classification as the reference.

Level II (1)[a]						Level II (2)[b]				
Predictor	κ	Accuracy	p	Sensitivity	Specificity	Predictor	κ	Accuracy	p	Sensiti
			Top five							
MoCA (23)	0.62	0.94	.017	0.50	0.99	MoCA (12)	0.85	0.99	.028	0.86
MoCA (3)	0.62	0.94	.017	0.50	0.99	MoCA (16)	0.85	0.99	.028	0.86
MoCA (51)	0.69	0.94	.017	0.70	0.97	MoCA (2)	0.83	0.99	.028	0.71
MoCA (11)	0.61	0.94	.032	0.55	0.98	MoCA (20)	0.85	0.99	.028	0.86
MoCA (13)	0.67	0.94	.032	0.70	0.97	MoCA (4)	0.83	0.99	.028	0.71
			Bottom five							
MoCA (5)	0.59	0.93	.093	0.60	0.97	MoCA (50)	0.70	0.98	.168	0.71
sMoCA (1)	0.69	0.93	.093	0.95	0.93	MoCA (54)	0.70	0.98	.168	0.71
MMSE (3)	0.43	0.93	.136	0.32	0.99	MoCA (58)	0.70	0.98	.168	0.71
MoCA (25)	0.55	0.92	.143	0.55	0.97	MMSE (2)	0.66	0.98	.168	0.57
MoCA (39)	0.53	0.92	.143	0.50	0.97	MMSE (4)	0.53	0.97	.296	0.43

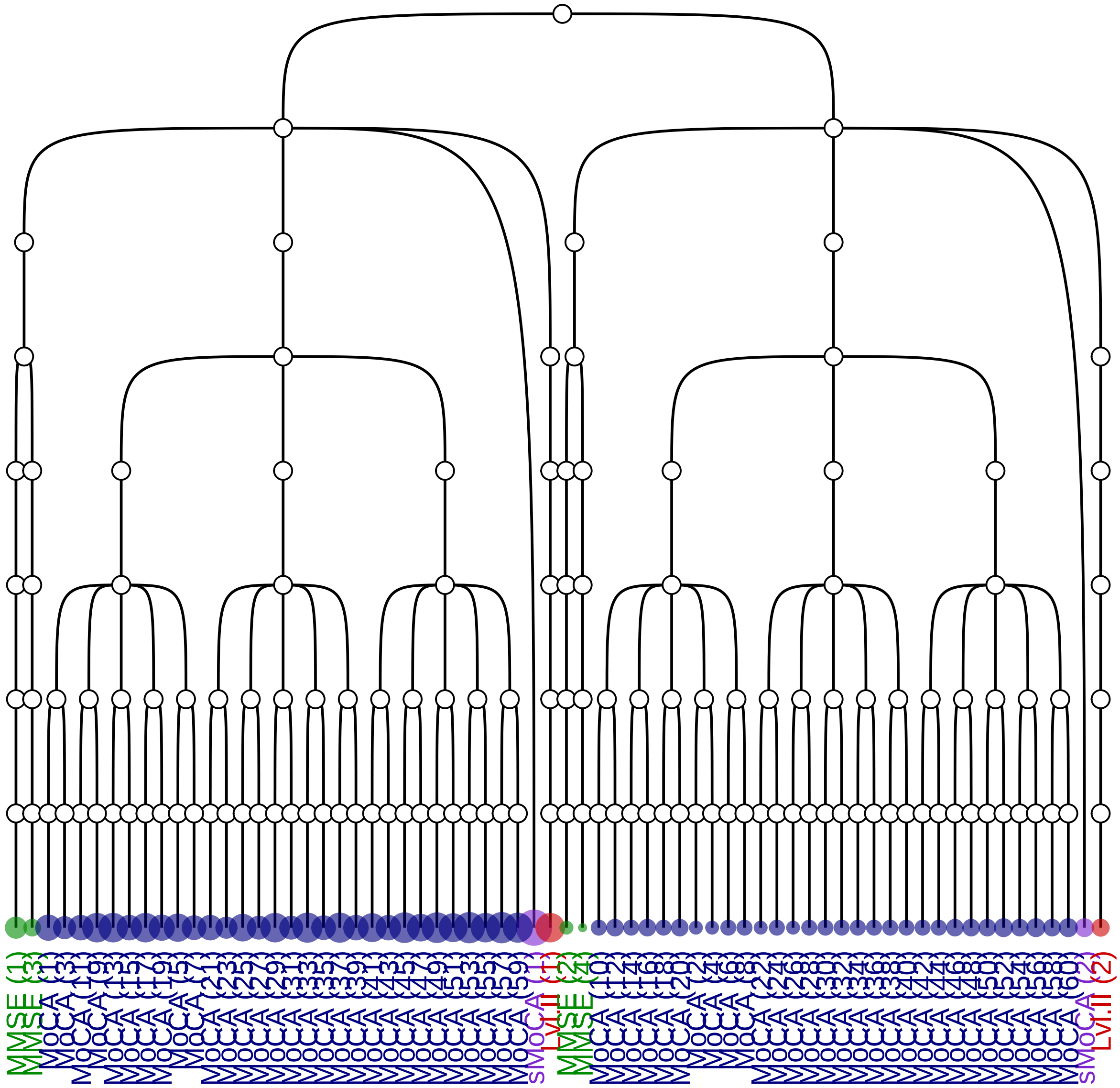
[a] IADL deficite was defined as FAQ (total score) > 7

[b] IADL deficite was defined as FAQ (item 9) > 1

κ: Cohen's κ; p: p-value associated with a one-sided Exact Binomial Test comparing the Accuracy to the No Information Rate; The table shows five most accurate (Top five) and five least accurate (Bottom five) Level I algrithms for Parkinson's Disease Dementia (PDD) in predicting Level II classficiation of PDD. The algorithms were grouped by their definition of the deficit in Instrumental Activities of Daily Living (IADLs). The items comprising each listed algorithm can be found in Table A1.

Figure 1

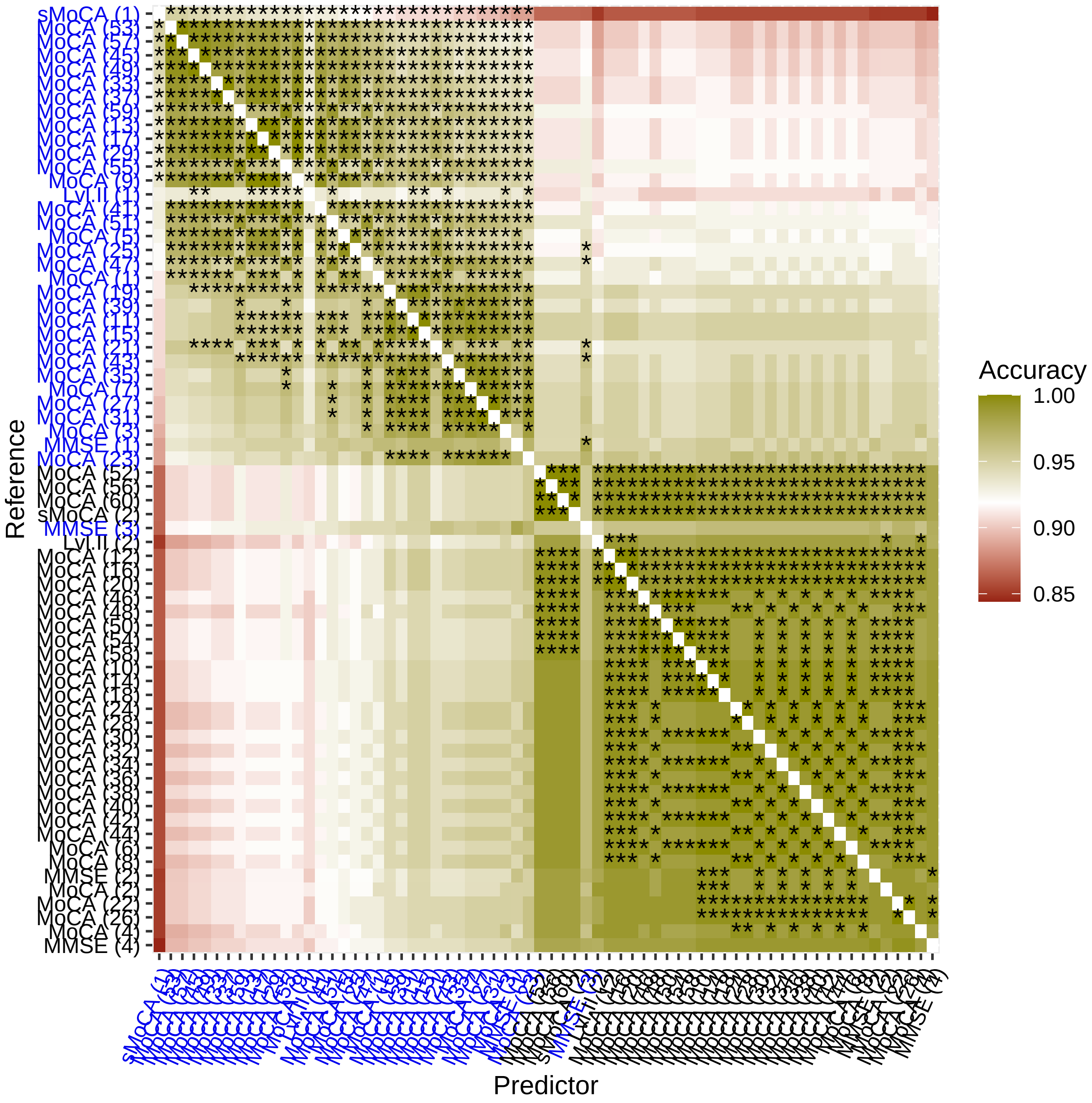
A dendrogram representing algorithms for probable Parkinson's Disease Dementia (PDD) construction process.



Note. The dendrogram illustrates the decision process used to construct algorithms for probable Parkinson's Disease Dementia (PDD). The second level depicts the definition of instrumental activities of daily living (IADL) deficit (FAQ total > 7 on the left, FAQ item 9 > 1 on the right). The third level indicates the selection of the screening instrument (MMSE, MoCA, sMoCA, or none in the case of Level II). Lower branches represent the selection of neuropsychological tests used to define cognitive impairment in executive function, attention, memory, and language, ordered from top to bottom as depicted in the dendrogram. Algorithms based on the MMSE are shown in green, those based on the MoCA in blue, on the sMoCA in purple, and Level II algorithms in red.

Figure 2

Prediction accuracy matrix.



Note. The matrix depicts classification accuracy of algorithms for PDD depicted on x-axis in predicting outcomes based on algorithms on the y-axis. Algorithms printed in blue defined IADL deficit by FAQ total score, algorithms printed in black defined IADL deficit by FAQ item 9 response. Cases with asterisk indicate predictive accuracy statistically significantly higher than the No Information Rate.

Appendix

Derivation of the Algorithms Set

Both, the original PDD criteria (Dubois et al., 2007) and the call for their change (Kulisevsky et al., 2024) allow for several distinct combinations of items to be used to define cognitive impairment. Consequently, in this study we derived all algorithms for probable PDD on Level I that are in line with published criteria. This procedure parallel the diagnostic algorithm outlined in Table 2 of Dubois et al. (2007). Specifically, in this study, we varied the exact specification of items 3-5 of this table (i.e., the measure of global cognitive impairment, the measure of the impact on IADLs and the measures of impaired cognition).

For each set of criteria (MMSE-based, MoCA-based, sMoCA-based and Level II), we first specified the items and then the thresholds for each item used to define probable PDD. If more than one option was present in either the choice of the item or the choice of the threshold, we created an algorithm for each choice in turn. The final set of algorithms was arrived at by computing the Cartesian product of all possibilities provided by varying items and thresholds. All combinations are presented in Table A1.

For MMSE-based algorithms, the following sets of items served as the basis:

$Global = \{MMSE < 26\}$

$Attention = \{Sevens\ backwards < 4\}$

$Executive = \{Clock\ drawing < 2, Lexical\ fluency\ (S) < 10\}$

$Construction = \{Pentagons < 1\}$

$Memory = \{Three\text{-}words\ recall < 3\}$

$IADL = \{FAQ > 7, FAQ\ (it.9) > 1\}$

The ensuing Cartesian product $Global \times Attention \times Executive \times Construction \times Memory \times IADL$ results in $1 \times 1 \times 2 \times 1 \times 1 \times 2 = 4$ MMSE-based algorithms for probable PDD.

For MoCA-based algorithms, the following sets of items served as the basis:

$Global = \{MoCA < 26\}$

$Attention = \{Sevens\ backwards < 3\}$

$Executive = \{Clock\ drawing < 2, Clock\ drawing < 3, Lexical\ fluency\ (K) < 11\}$

$Construction = \{Cube\ drawing < 1\}$

$Memory = \{Five\text{-}words\ recall < 1, Five\text{-}words\ recall < 2, Five\text{-}words\ recall < 3, Five\text{-}words\ recall < 4, Five\text{-}words\ recall < 5\}$

$Language = \{Abstraction < 2, Animal\ naming < 3\}$

$IADL = \{FAQ > 7, FAQ\ (it.9) > 1\}$

Note that the additional language domain adds complexity to establishing a diagnostic algorithm because simply by adding it to the set of items, the number of potential algorithms doubles. Further complexity is added by the fact that there are so far no guidelines for selecting a diagnostic threshold for Clock drawing and Five-words recall tests, both of which differ from their counterparts used by Dubois et al. (2007). Finally, although the Sevens backwards item has different thresholds in MoCA-based compared to MMSE-based algorithms, this difference is solely due to a difference in scoring whereby 3 points in MoCA correspond to 4 or 5 points in MMSE. The Seven backwards item threshold for MoCA-based algorithms used in this study is thus equivalent to its MMSE-based counterpart.

Computing the Cartesian product $Global \times Attention \times Executive \times Construction \times Memory \times Language \times IADL$ yields $1 \times 1 \times 3 \times 1 \times 5 \times 2 \times 2 = 60$ distinct MoCA-based algorithms for probable PDD.

For sMoCA-based algorithms, the following sets of items served as the basis:

$Global = \{sMoCA < 13\}$

$IADL = \{FAQ > 7, FAQ\ (it.9) > 1\}$

yielding $Global \times IADL$, i.e., $1 \times 2 = 2$ distinct sMoCA-based algorithms for probable PDD.

Finally, the Level II algorithms were based on the following sets of items:

$Attention = \{z(TMT\ A) < -1.5\ \cup z(WAIS\ DSB) < -1.5\}$

$Executive = \{z(CF\ A) < -1.5\ \cup z(PST\ C) < -1.5\}$

$Construction = \{z(JoLO) < -1.5\ \cup z(CLOXI) < -1.5\}$

$Memory = \{z(RAVLT\ DR) < -1.5\ \cup z(BVMTR\ DR) < -1.5\ \cup z(WMS\text{-}III\ Family\ Pictures) < -1.5\}$

$Language = \{z(WAIS\ Similarities) < -1.5\ \cup z(BNT\ 60) < -1.5\}$

$IADL = \{FAQ > 7, FAQ\ (it.9) > 1\}$

where $z()$ denotes calculation of age, sex and education adjusted z-score. This yields $1 \times 1 \times 1 \times 1 \times 1 \times 2 = 2$ distinct Level II algorithms for probable PDD in the current study. All but the BNT 60 item were evaluated using regression norms published by Bezdicek et al. (2017). Since the original article used BNT 30 instead of BNT 60, we approximated the deficit in BNT 60 by comparing patients' raw score to age- and education-specific normative values reported by Zemanová et al. (2016). Specifically, patients whose BNT 60 score fell below 5th percentile of their demographic group in Table 6 of Zemanová et al. (2016) were considered to show signs of impaired performance.

Operationalization of Impaired Cognition

In the original criteria, item 4 of Level I criteria, i.e., impaired cognition, was defined as follows: *"The proposed diagnostic criteria require a profile of cognitive deficits, typical of those described for PD-D, in two or more of four domains."* (Dubois et al., 2007, p. 2316) Consequently, we defined impaired cognition as a deficit in two or more domains of four in MMSE-based criteria and as a deficit in two or more of five domains in MoCA-based criteria. sMoCA-based criteria omitted the "impaired cognition" item altogether because they were intended as a shorter screening alternative to classical Level I assessment. Finally, for the Level II criteria, we employed standard definition of impaired cognition as the *"[i]mpairment on at least two neuropsychological tests, represented by either two impaired tests in one cognitive domain or one impaired test in two different cognitive domains."* (Litvan et al., 2012, Table 1)

Theoretical and Empirical Estimands

In this study, we follow the framework proposed by Lundberg et al. (2021) for specifying targets of inference (i.e., the estimands) in qunatitative sciences to increase transparency and connect statistical evidence to relevant theory. Table A2 contains verbal description of the components relating to each of our proclaimed research objectives and map them to the population quantity of interest (the theoretical estimand), data-dependent quantity that could be estimated (the empirical estimand) and quantities that are reported in the study (statistical estimates).

The RO1 - to estimate the PDD rate and evaluate the diagnostic variability and concordance across different algorithms of probable PDD - was divided into four distinct research objectives:

- to estimate the rate of PDD within PD (RO1.1),
- to estimated variability of this rate (RO1.2),
- to evaluate predictive value of demographic variables for probable PDD classification (RO1.3) and
- to evaluate concordance between different probable PDD operationalizations and criteria (RO1.4).

Estimates relating to RO1.1 and RO1.3 cannot be safely generalized beyond a population of PD patients that are candidates for DBS due to the systematic differences between DBS candidates pool and general PD population (such as the lower age of DBS candidates compared to the general PD population). On the other hand, the estimates relating to RO1.4 (and to a lesser degree to RO1.2[2]) may not be substantially influenced by the sample at hand as the primary

[2] Because the quantity of interest is a rate and could thus be though of as a sum of binomially distributed PDD occurences divided by the total number of patients, its variance will likely systematically vary with its mean. Specifically, as the rate goes from extremes to 0.5, the variance increases. Consequently, if our estimate of the rate was lower than the true population rate, e.g., because our sample includes younger patients compared to the general PD population, our estimate of variance would also be lower than the true variance of PDD rate in the general PD

source of their variance might come from variability in measures employed (e.g., MMSE vs MoCA to assess global cognitive performance) rather than variability in patients’ performance. Assuming that there is no substantial Differential Item Functioning for DBS candidates compared to a broader population of patients with PD, the estimates relating to RO1.4 can be cautiously generalized beyond the current sample.

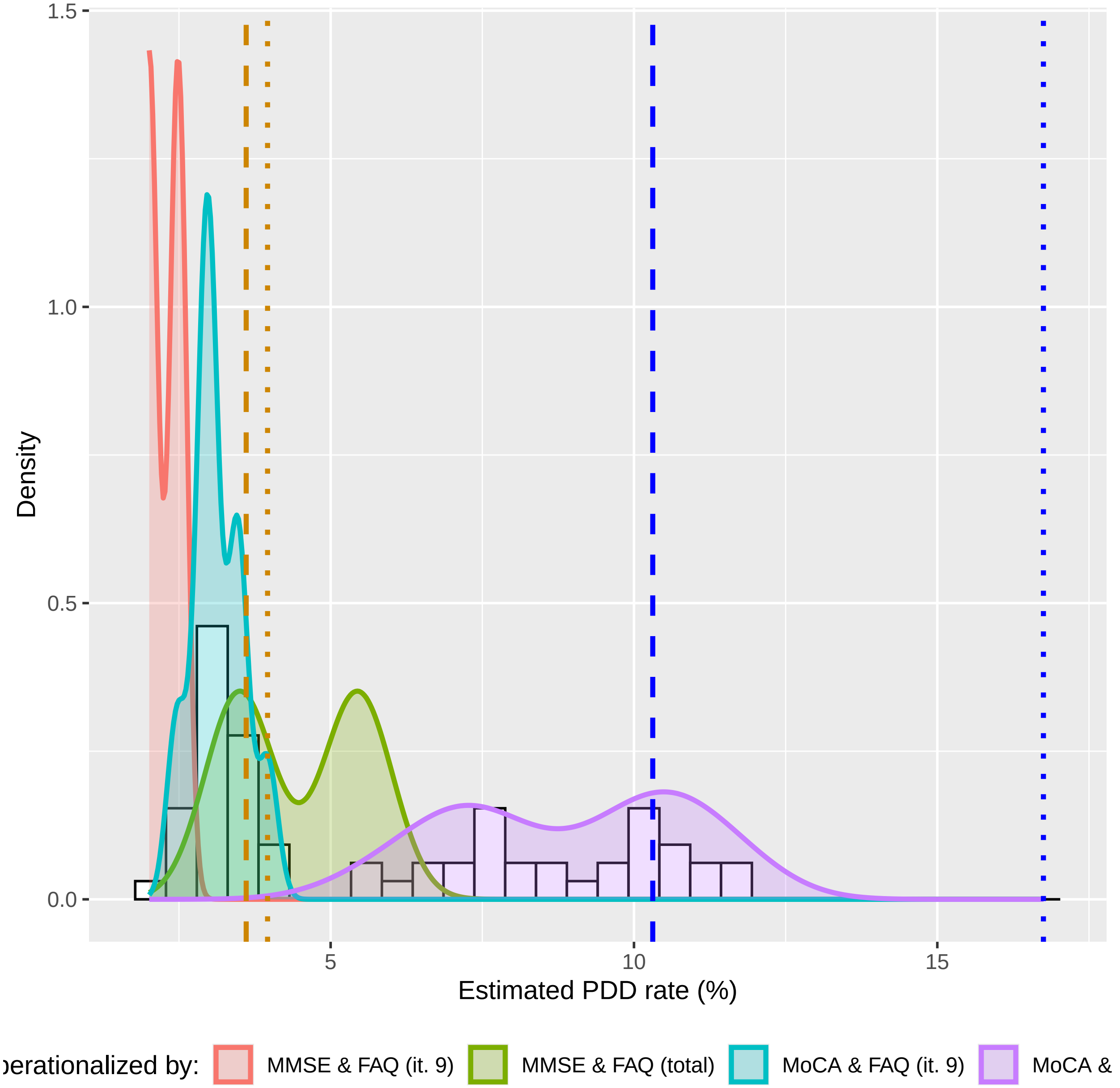
Finally, for the RO2, the theoretical estimand is defined as the set of diagnostic components whose variation systematically alters the probability of a probable PDD diagnosis. This aspect of the study is exploratory in nature. Empirically, we assess the contribution of each diagnostic feature by examining how variations in operational definitions (e.g., domain-specific thresholds, criteria for functional impairment) influence the statistical estimates derived for the first objective. This allows us to identify the diagnostic elements most responsible for between-algorithm discrepancies.

Supplementary Presentation of Results

population. Nonetheless, the between-algorithm variability may not be affected by this phenomenon as unlike variability of PDD rate, we do not have reason to assume it comes about by summing independent binomial events.

Figure A1

Summary of the estimates of probable PDD rate.



Note. Vertical lines represent estimates arrived at by using sMoCA (dotted) or Level II (dashed) with FAQ item 9 (orange) or FAQ total score (blue) as criteria for probable PDD. The percentages were calculated from all available cases (see Table A3 for numerical summary).

Table A1

Summary of all algortihms for probable PDD used in the study.

Algorithm	Global deficit	Attention	Executive function	
Lvl.II (1)	-	TMT A < -1.5 OR WAIS DS < -1.5	CF A < -1.5 OR PST C < -1.5	Jo
Lvl.II (2)	-	TMT A < -1.5 OR WAIS DS < -1.5	CF A < -1.5 OR PST C < -1.5	Jo
MMSE (1)	Total score < 26	Sevens < 4	Clock Drawing < 2	
MMSE (2)	Total score < 26	Sevens < 4	Clock Drawing < 2	
MMSE (3)	Total score < 26	Sevens < 4	VF S < 10	
MMSE (4)	Total score < 26	Sevens < 4	VF S < 10	
MoCA (1)	Total score < 26	Sevens < 3	Clock drawing < 2	
MoCA (10)	Total score < 26	Sevens < 3	Clock drawing < 2	
MoCA (11)	Total score < 26	Sevens < 3	Clock drawing < 2	
MoCA (12)	Total score < 26	Sevens < 3	Clock drawing < 2	
MoCA (13)	Total score < 26	Sevens < 3	Clock drawing < 2	
MoCA (14)	Total score < 26	Sevens < 3	Clock drawing < 2	
MoCA (15)	Total score < 26	Sevens < 3	Clock drawing < 2	
MoCA (16)	Total score < 26	Sevens < 3	Clock drawing < 2	
MoCA (17)	Total score < 26	Sevens < 3	Clock drawing < 2	
MoCA (18)	Total score < 26	Sevens < 3	Clock drawing < 2	
MoCA (19)	Total score < 26	Sevens < 3	Clock drawing < 2	
MoCA (2)	Total score < 26	Sevens < 3	Clock drawing < 2	
MoCA (20)	Total score < 26	Sevens < 3	Clock drawing < 2	
MoCA (21)	Total score < 26	Sevens < 3	VF K < 11	
MoCA (22)	Total score < 26	Sevens < 3	VF K < 11	
MoCA (23)	Total score < 26	Sevens < 3	VF K < 11	
MoCA (24)	Total score < 26	Sevens < 3	VF K < 11	
MoCA (25)	Total score < 26	Sevens < 3	VF K < 11	
MoCA (26)	Total score < 26	Sevens < 3	VF K < 11	

Table A2

Mapping between research objectives and quantities of interest in the current study.

To estimate the rate of PDD within PD.
To assess variability of PDD diagnosis depending on the algorithm applied.
To evaluate predictive information provided by demographic variables for probable PDD diagnosis.
To evaluate the diagnostic concordance between different PDD algorithms within and between PDD criteria.
To identify algorithms’ components that contribute to variability in probable PDD diagnosis within and across cr

Table A3

Estimates of the rate of probable PDD in the sample.

Algorithm	N	Rate
sMoCA (1)	203	34 (16.75%)
MoCA (53)	203	24 (11.82%)
MoCA (57)	203	24 (11.82%)
MoCA (45)	203	23 (11.33%)
MoCA (49)	203	23 (11.33%)
MoCA (33)	203	22 (10.84%)
MoCA (37)	203	22 (10.84%)
MoCA (59)	203	22 (10.84%)
MoCA (13)	203	21 (10.34%)
MoCA (17)	203	21 (10.34%)
MoCA (29)	203	21 (10.34%)
MoCA (55)	203	21 (10.34%)
MoCA (9)	203	21 (10.34%)
Lvl.II (1)	194	20 (10.31%)
MoCA (41)	203	20 (9.85%)
MoCA (51)	203	20 (9.85%)
MoCA (5)	203	19 (9.36%)
MoCA (25)	203	18 (8.87%)
MoCA (47)	203	18 (8.87%)
MoCA (1)	203	16 (7.88%)
MoCA (19)	203	16 (7.88%)
MoCA (11)	203	15 (7.39%)
MoCA (15)	203	15 (7.39%)
MoCA (21)	203	15 (7.39%)
MoCA (39)	203	15 (7.39%)
MoCA (43)	203	15 (7.39%)

Table A4

Level I algorithms for probable PDD as predictors of Level II classification as the reference arranged by their balanced accuracy score.

Level II (1)[a]						Level II (2)[b]				
Predictor	κ	Accuracy	p	Sensitivity	Specificity	Predictor	κ	Accuracy	p	Sensiti
			Top five							
sMoCA (1)	0.69	0.93	.093	0.95	0.93	MoCA (12)	0.85	0.99	.028	0.86
MoCA (45)	0.68	0.94	.032	0.75	0.96	MoCA (16)	0.85	0.99	.028	0.86
MoCA (49)	0.68	0.94	.032	0.75	0.96	MoCA (20)	0.85	0.99	.028	0.86
MoCA (53)	0.66	0.93	.056	0.75	0.95	MoCA (52)	0.79	0.98	.078	0.86
MoCA (57)	0.66	0.93	.056	0.75	0.95	MoCA (56)	0.79	0.98	.078	0.86
			Bottom five							
MoCA (21)	0.55	0.93	.093	0.50	0.98	MoCA (58)	0.70	0.98	.168	0.71
MoCA (35)	0.55	0.93	.093	0.50	0.98	MoCA (22)	0.66	0.98	.168	0.57
MoCA (39)	0.53	0.92	.143	0.50	0.97	MoCA (26)	0.66	0.98	.168	0.57
MMSE (1)	0.55	0.93	.056	0.45	0.99	MMSE (2)	0.66	0.98	.168	0.57
MMSE (3)	0.43	0.93	.136	0.32	0.99	MMSE (4)	0.53	0.97	.296	0.43

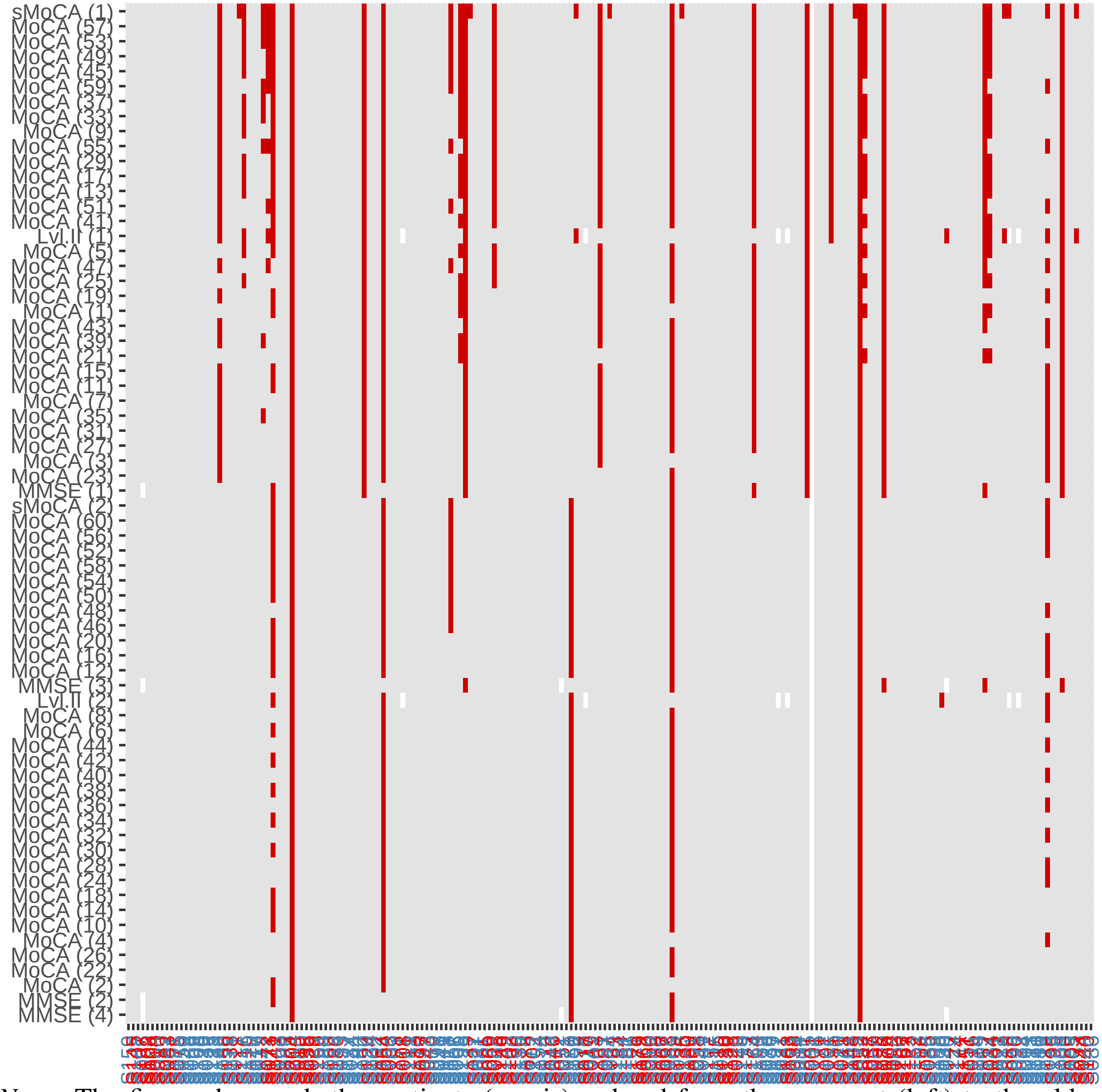
[a] IADL deficite was defined as FAQ (total score) > 7

[b] IADL deficite was defined as FAQ (item 9) > 1

κ: Cohen’s κ; p: p-value associated with a one-sided Exact Binomial Test comparing the Accuracy to the No Information Rate; The table shows five most accurate (Top five) and five least accurate (Bottom five) Level I algrithms for Parkinson’s Disease Dementia (PDD) in predicting Level II classficiation of PDD. The algorithms were grouped by their definition of the deficit in Instrumental Activities of Daily Living (IADLs). The items comprising each listed algorithm can be found in Table A1.

Figure A2

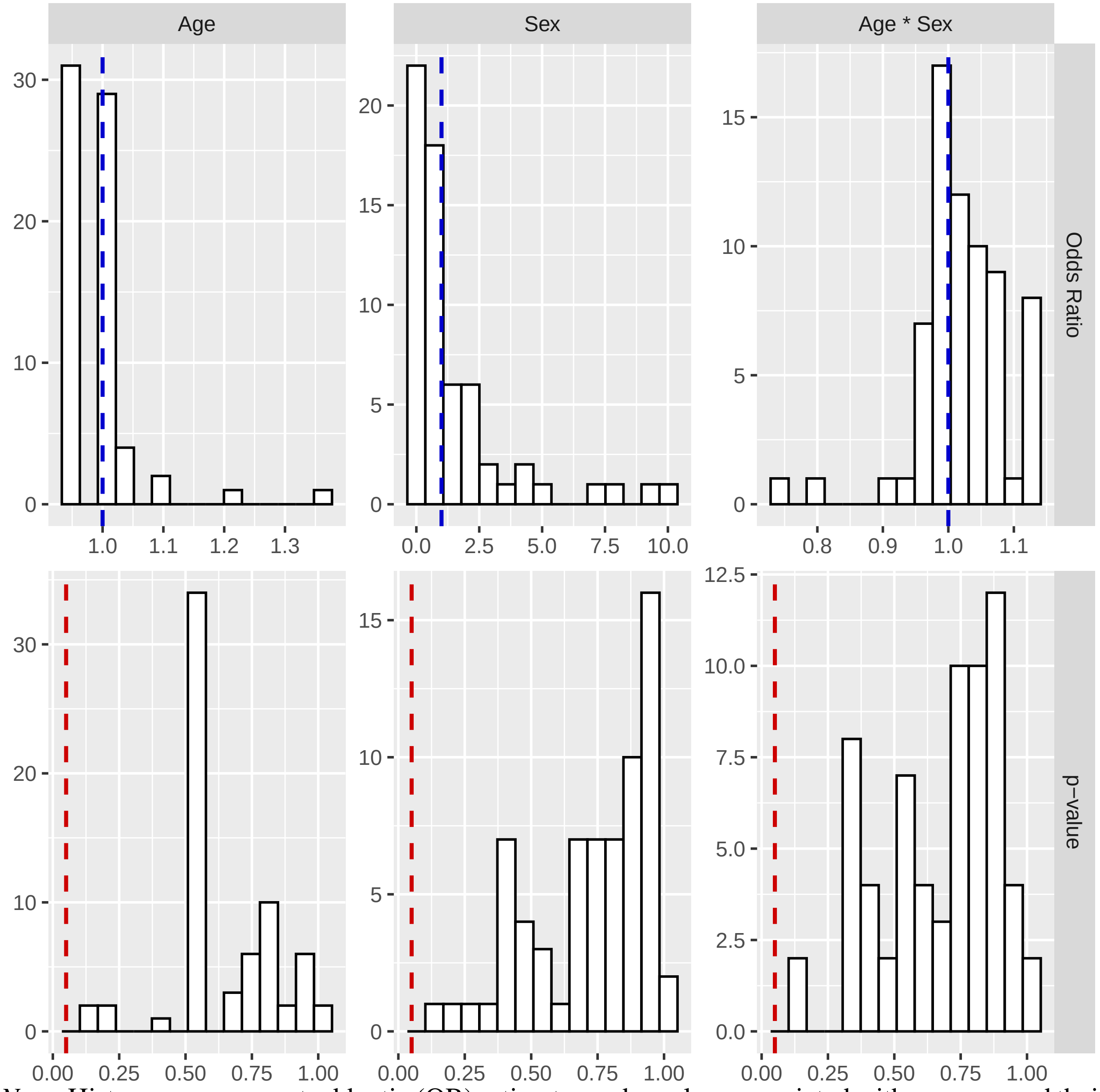
Representation of study data.



Note. The figure shows whether patients (x-axis) ordered from the youngest (left) to the oldest (right) were classified as probable PDD by each tested algorithm (y-axis) ordered from the one with the lowest (bottom) to the highest (top) PDD rate estimate. Patients printed in red are women, patients printed in blue are men. Red cells indicate probable PDD diagnosis, grey cells indicate non-PDD diagnosis and white cells indicate missing diagnosis.

Figure A3

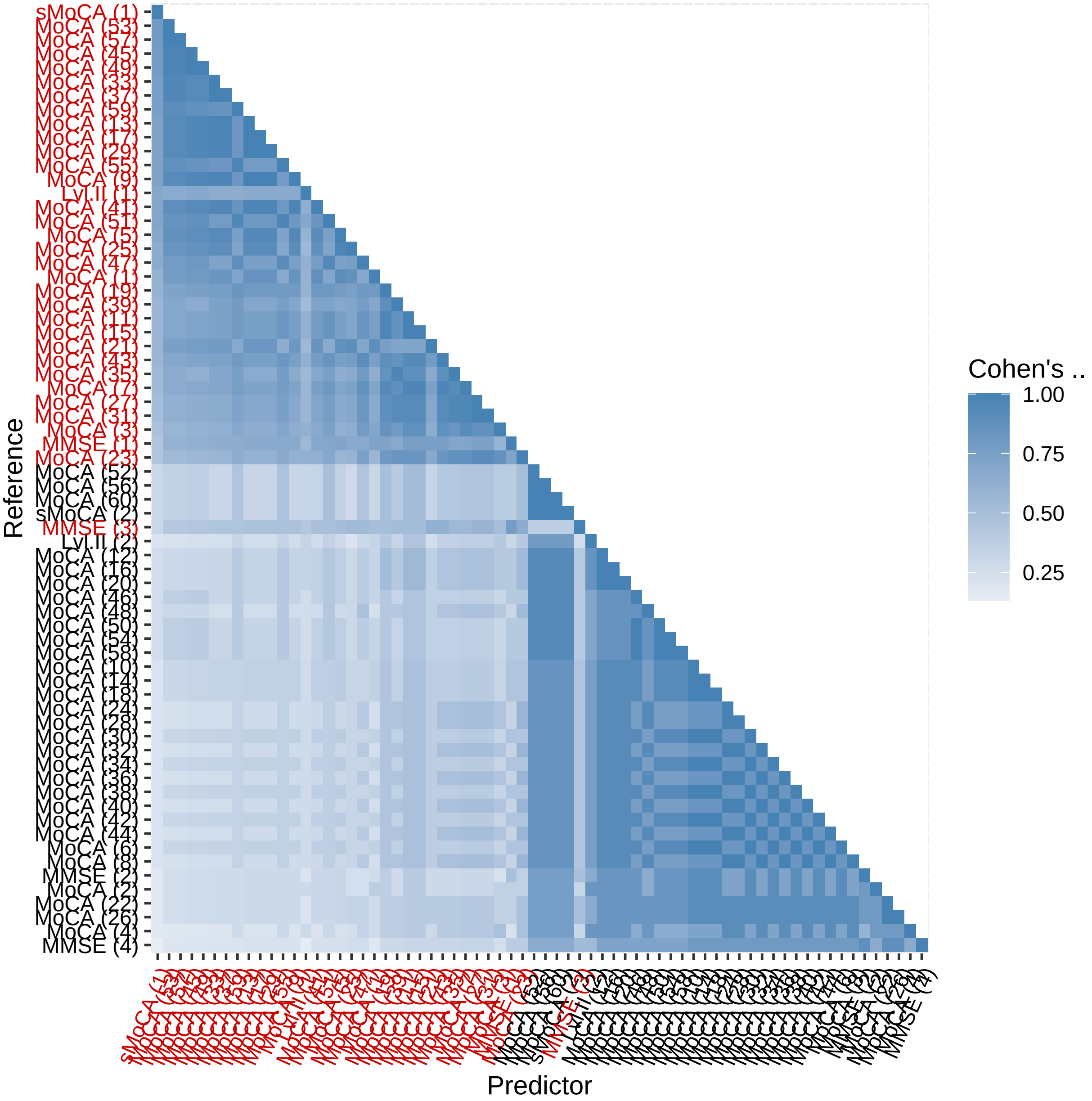
Summary of logistic regressions parameters prediction probable PDD by age and sex.



Note. Histograms represent odd ratio (OR) estimates and p-values associated with age, sex, and their interaction as predictors of each of the 68 probable PDD classification. In the case of parameters for sex, values higher than 20 were omitted for clarity. Vertical lines indicate OR = 1 and p = .05.

Figure A4

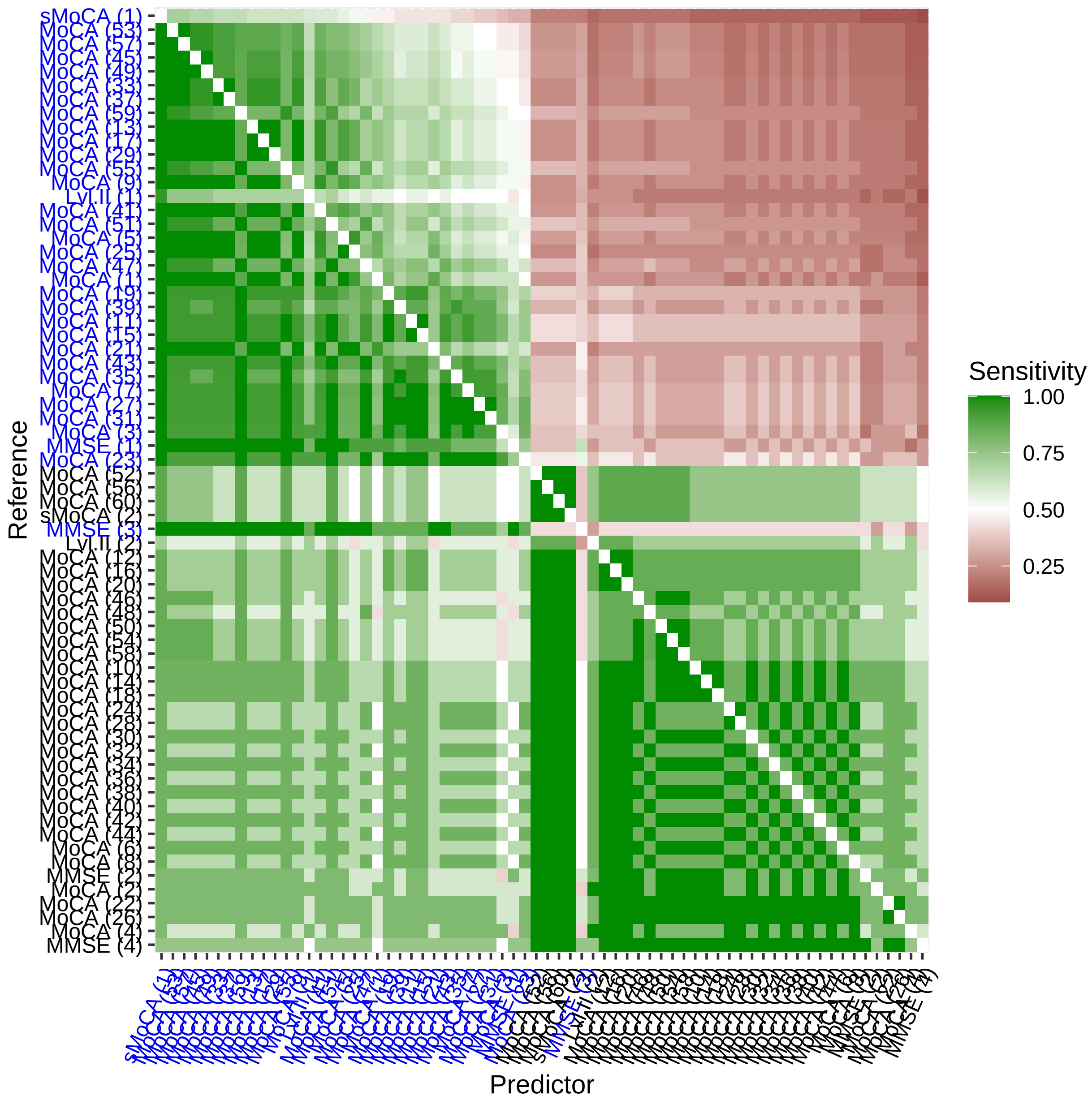
Cohen’s κ matrix.



Note. The matrix depicts Cohen’s κ measuring agreement between algorithms for PDD. Algorithms printed in red defined IADL deficit by FAQ total score, algorithms printed in black defined IADL deficit by FAQ item 9 response.

Figure A5

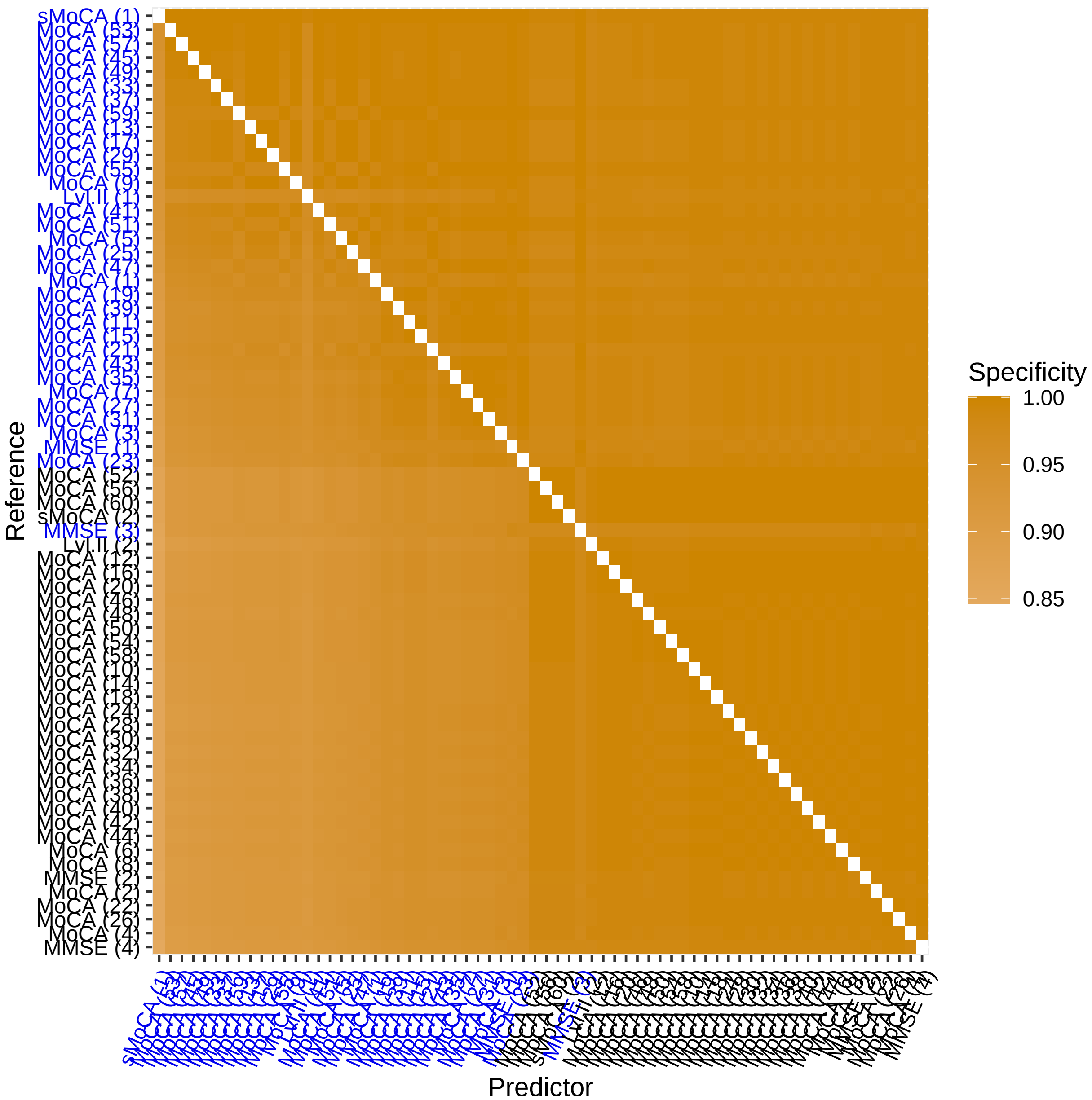
Sensitivity matrix.



Note. The matrix depicts sensitivity of algorithms for PDD depicted on x-axis in predicting outcomes based on algorithms on the y-axis. Algorithms printed in blue defined IADL deficit by FAQ total score, algorithms printed in black defined IADL deficit by FAQ item 9 response.

Figure A6

Specificity matrix.



Note. The matrix depicts specificity of algorithms for PDD depicted on x-axis in predicting outcomes based on algorithms on the y-axis. Algorithms printed in blue defined IADL deficit by FAQ total score, algorithms printed in black defined IADL deficit by FAQ item 9 response.