RESEARCH ARTICLE

Detecting Mild Cognitive Deficits in Parkinson's Disease: Comparison of Neuropsychological Tests

Jeroen Hoogland, MD,[1*] Lennard L. van Wanrooij,[1] Judith A. Boel, MSc,[1] Jennifer G. Goldman, MD, MS,[2] Glenn T. Stebbins, PhD,[2] John C. Dalrymple-Alford, PhD,[3] Connie Marras, MD, PhD,[4] Charles H. Adler, MD, PhD,[5] Carme Junque, PhD,[6] Kenn F. Pedersen, MD, PhD,[7] Brit Mollenhauer, MD,[8] Cyrus P. Zabetian, MD, MS,[9] Paul J. Eslinger, PhD,[10] Simon J. G. Lewis, MD,[11] Ruey-Meei Wu, PhD,[12] Martin Klein, PhD,[13] Maria C. Rodriguez-Oroz, MD, PhD,[14] Davide M. Cammisuli, PhD,[15] Paolo Barone, MD, PhD,[16] Roberta Biundo, PhD,[17] Rob M. A. de Bie, MD, PhD,[1] Ben A. Schmand, PhD,[18,19] Alexander I. Tröster, PhD,[20] David J. Burn, MD, PhD,[21] Irene Litvan, MD, PhD,[22] J. Vincent Filoteo, PhD,[22,23] Gert J. Geurtsen, PhD,[18] and Daniel Weintraub, MD,[24] on behalf of the IPMDS Study Group "Validation of Mild Cognitive Impairment in Parkinson Disease"

[1]*Department of Neurology, Academic Medical Center Amsterdam, The Netherlands*
[2]*Department of Neurological Sciences, Section of Parkinson Disease and Movement Disorders, Rush University Medical Center, Chicago, Illinois, USA*
[3]*New Zealand Brain Research Institute, Brain Research New Zealand — Rangahau Roro Aotearoa, Christchurch, New Zealand*
[4]*Morton and Gloria Shulman Movement Disorders Clinic and the Edmond J Safra Program in Parkinson's disease, Toronto Western Hospital, University of Toronto, Toronto, Ontario, Canada*
[5]*Arizona Parkinson's Disease Consortium, Mayo Clinic Arizona, Scottsdale, Arizona, USA and Banner Sun Health Research Institute, Sun City, Arizona, USA*
[6]*Department of Medicine, Faculty of Medicine, IDIBAPS, University of Barcelona, Barcelona, Spain*
[7]*The Norwegian Centre for Movement Disorders, Department of Neurology, and Memory Clinic, Stavanger University Hospital, Stavanger, Norway*
[8]*Paracelsus-Elena-Klinik, Kassel, Germany, and University Medical Center Goettingen, Department of Neurology, Goettingen, Germany*
[9]*VA Puget Sound Health Care System and Department of Neurology, University of Washington School of Medicine, Seattle, Washington, USA*
[10]*Department of Neurology, Penn State Hershey Medical Center, Hershey, Pennsylvania, USA*
[11]*Brain & Mind Centre, The University of Sydney, Sydney, Australia*
[12]*Department of Neurology, National Taiwan University Hospital, College of Medicine, National Taiwan University, Taipei, Taiwan*
[13]*Department of Medical Psychology, section Medical Neuropsychology, VU University Medical Center, Amsterdam, The Netherlands*
[14]*Department of Neurology, Hospital Donostia, Donostia, San Sebastian and Ikerbasque, Basque Foundation for Science, Bilbao, Spain*
[15]*Fifth Local Sanitary Unit, Hospital Psychology, Pisa, Italy and Department of Surgical, Medical, Molecular, and Critical Area Pathology, Pisa University School of Medicine, Pisa, Italy*
[16]*Center for Neurodegenerative Diseases (CEMAND) Neuroscience Section, Department of Medicine University of Salerno, Salerno, Italy*
[17]*San Camillo Hospital IRCCS, Venice, Italy*
[18]*Department of Medical Psychology, Academic Medical Center, Amsterdam, The Netherlands*
[19]*Department of Psychology, University of Amsterdam, The Netherlands*
[20]*Department of Clinical Neuropsychology and Center for Neuromodulation, Barrow Neurological Institute, Phoenix, Arizona, USA*
[21]*Institute of Neuroscience, Newcastle University, Newcastle upon Tyne, UK*
[22]*Department of Neurosciences University of California San Diego, Parkinson and Other Movement Disorders Center, San Diego, California, USA*
[23]*Department of Psychiatry, University of California San Diego, and VA San Diego Healthcare System, San Diego, California, USA*
[24]*Departments of Psychiatry and Neurology, University of Pennsylvania School of Medicine, and Parkinson's Disease and Mental Illness Research, Philadelphia Veterans Affairs Medical Center, Philadelphia, Pennsylvania, USA*

*__Correspondence to__: Jeroen Hoogland, MD, Meibergdreef 9, 1105AZ, Amsterdam, the Netherlands; j.hoogland@amc.nl

IPMDS Study Group "Validation of Mild Cognitive Impairment in Parkinson Disease" (see Appendix)

Gert J. Geurtsen and Daniel Weintraub MD, senior authors, contributed equally to this article.

Relevant conflicts of interest/financial disclosures: .

Funding agencies: Study funded by the Michael J. Fox Foundation for Parkinson's Research and Dutch Parkinson Disease Association.

Received: 15 February 2018; **Revised:** 11 May 2018; **Accepted:** 25 June 2018

Published online 00 Month 2018 in Wiley Online Library (wileyonlinelibrary.com). DOI: 10.1002/mds.110

ABSTRACT: Background: Numerous neuropsychological tests and test versions are used in Parkinson's disease research, but their relative capacity to detect mild cognitive deficits and their comparability across studies are unknown. The objective of this study was to identify neuropsychological tests that consistently detect cognitive decline in PD across studies. **Methods:** Data from 30 normed neuropsychological tests across 20 international studies in up to 2908 nondemented PD patients were analyzed. A subset of 17 tests was administered to up to 1247 healthy controls. A 2-step meta-analytic approach using standardized scores compared performance in PD with normative data. **Results:** Pooled estimates of the differences between PD and site-specific healthy controls identified significant cognitive deficits in PD patients on 14 test scores across 5 commonly assessed cognitive domains (attention or working memory, executive, language, memory, and visuospatial abilities), but healthy control performance was statistically above average on 7 of these tests. Analyses based on published norms only, as opposed to direct assessment of healthy controls, showed high between-study variability that could not be accounted for and led to inconclusive results. **Conclusions:** Normed neuropsychological tests across multiple cognitive domains consistently detect cognitive deficits in PD when compared with site-specific healthy control performance, but relative PD performance was significantly affected by the inclusion and type of healthy controls versus the use of published norms only. Additional research is needed to identify a cognitive battery that can be administered in multisite international studies and that is sensitive to cognitive decline, responsive to therapeutic interventions, and superior to individual cognitive tests.

Key Words: Parkinson disease; MCI; mild cognitive impairment; cognition; neuropsychological

Progression to dementia in Parkinson's disease (PDD) is increasingly recognized as a common and disabling feature of the disease.[1–3] Initial intervention studies focused on patients with PDD or the related dementia with Lewy bodies (DLB),[4–7] but there is also growing interest in the potential for interventions prior to dementia, such as at the stage of mild cognitive impairment (PD-MCI). A diagnosis of PD-MCI is a strong risk factor for progression to PDD.[8–12] Therefore, there is interest in the epidemiology, types of cognitive impairment, and optimal assessment and diagnosis of PD-MCI.[13]

In nondemented PD patients it is recommended by the International Parkinson and Movement Disorder Society (IPMDS) PD-MCI criteria that a battery of neuropsychological tests should test abilities across 5 cognitive domains (ie, attention/working memory, executive function, language, memory, and visuospatial function), either in a preferred comprehensive (level II) or abbreviated (level I) fashion.[14] However, the specific neuropsychological tests used to ascertain a PD-MCI diagnosis across studies remain highly varied in content and psychometric properties. Ideally these tests should demonstrate good test characteristics, including sensitivity to the presence of the earliest stage of cognitive decline, progression of decline, and improvement with cognition-enhancing treatment; be associated with other clinically-important outcomes (eg, functional abilities and clinical global impression); be correlated with neurobiological markers of cognitive impairment; and be readily accessible and relatively simple to administer in persons with a wide range of disease, cultural, and demographic characteristics.

Establishing a core neuropsychological battery demonstrated to be sensitive to cognitive deficits and decline in nondemented PD could improve clinical research by encouraging the use of validated, standardized protocols in cross-sectional and longitudinal research, a step beyond the expert consensus recommendations for cognitive testing by the National Institute of Neurological Disorders and Stroke (NINDS)-funded Udall Centers[15] and the NINDS Common Data Element Project.[16] The primary aim of this study was to compare the relative sensitivity of neuropsychological tests used in our international consortium to detect cognitive deficits in nondemented PD patients across studies and provide a first step in the development of a core neuropsychological battery for use in PD clinical research studies focused on cognition. We used cognitive data from a large number of PD patients (more than 3000 nondemented patients) and healthy controls (more than 1000 persons) across multiple, international studies through the work of the IPMDS PD-MCI Validation Study Group.[17]

Methods

Study Design and Participants

The IPMDS PD-MCI Validation Study Group is an international consortium that was formed to validate the IPMDS PD-MCI criteria[14] in a pooled group of large extant databases from studies of PD cognition[9,17] (see Supplementary Co-Investigator File detailing the IPMDS Study Group "Validation of Mild Cognitive Impairment in Parkinson Disease").

Data from the consortium include clinical and neuropsychological measures from a total of 24 cohort studies in PD with a focus on cognition, 13 of these studies also including data on healthy controls (HCs). Supplementary

Table 1 and Table 1 show the characteristics of the included cohorts. Neuropsychological tests at each site were administered and scored by trained research personnel and in that country's official language.

To achieve our goals, a 2-step data-inclusion process was used. First, cross-sectional baseline data on neuropsychological tests were selected from the IPMDS PD-MCI consortium database when (1) normative scores based on published norms were available, to enable comparison of neuropsychological tests to each other and across slightly different versions sometimes used in different countries, and (2) the neuropsychological tests were administered in at least 2 countries and across at least 3 studies to increase the generalizability of our results. Details on the neuropsychological tests included are shown in Table 2.

Second, eligible patients (PD) and HCs were selected from the identified studies to establish a nondemented study cohort. Patients diagnosed with PDD at baseline visit were excluded; this classification was performed according to the diagnostic criteria specified in Supplementary Table 1, and when formal diagnostic criteria were not applied, patients were excluded as having possible dementia based on scoring below recommended cutoffs on available global cognitive screening instruments[18] as follows: Montreal Cognitive Assessment (MoCA; cutoff 20/21),[19,20] Mattis Dementia Rating Scale (MDRS; cutoff 129/130, a blend between the original recommended cutoff[21] and more recent recommendation),[21,22] and Mini-Mental State Examination (MMSE; cutoff 23/24).[23] The same cutoffs on the global cognitive measures were applied to HCs. To focus on typical PD patients, patients with a disease duration > 25 years were excluded. Applying these criteria led to the exclusion of an additional 72 participants (49 PD patients, 23 HCs), and another 46 participants (36 PD patients, 10 HCs) who had neither dementia status coded nor available cognitive screening tests (Fig. 1).

Statistical Analyses

Preprocessing of Neuropsychological Data

Normative neuropsychological scores were available as either *z* scores, *t* scores, or various types of scaled scores (eg, Wechsler scaled scores). The scores were transformed to a *z*-score scale based on their theoretical distributions. Furthermore, to avoid influential outliers, scores were winsorized to the -3.0 to 3.0 standard deviation (SD) range, which affected 2% of the data.

Incomplete Data

For quality control and to avoid incomplete data in large undefined subsets that are possibly not missing at random, data for any given test-cohort combination were not used if the test had >25% missing values within a PD cohort. Detailed study of reasons for missing data within the original studies was not feasible. In all other cases of missing data, multiple imputation was applied using the *mice* package in R.[24] Multiple imputation accounts for the relationship between variables, which is especially useful for correlated data (eg, neuropsychological test results) and takes into account uncertainty regarding incomplete data.[25] Incomplete data were assumed to be missing at random.[26] To allow for between-study heterogeneity in the imputation model, multiple imputation was applied within each cohort's data set.[27] Twenty imputations were run using predictive mean matching based on a model that included age, sex, years of education, Hoehn and Yahr stage, Unified Parkinson's Disease Rating Scale (UPDRS) part III score, disease duration, and neuropsychological measures, with PD-related measures contributing to imputations in PD patients only. Therefore, different imputation models were run for analyses A and B (described below) because of their different requirements. Summary measures were derived within each original study. Analyses were pooled over imputations using Rubin's rules,[28] which provides a means to combine within- and between-imputation variability into a single estimate.

Analysis A: Neuropsychological Performance in PD Using Published Norms

For analysis A, each test (ie, subtest score) was analyzed separately (see Table 2 for sample size for each test and published norms used). To evaluate whether PD patients deviated from normal performance on each of the available neuropsychological measures, their performance in terms of normative *z* scores was compared relative to the expected mean under normal performance (ie, zero). Exclusion of zero from the 95% confidence interval indicated rejection of the null hypothesis of no impairment. A 2-step meta-analysis approach was used to derive these estimates, using the pooled within-study estimates for both fixed and random effects inverse variance-weighted models.[29] The Hartung Knapp method, as implemented in R's *meta* package,[30] was used; this is recommended when the number of contributing studies is low.[31] A lower bound of the 95% confidence interval of I^2 over 50% indicated that greater than half the observed variability in the estimate was a result of between-study differences and was interpreted as a sign of large heterogeneity in performance.

In addition, the residual influence of age, sex, and education on the normative scores of PD patients was evaluated using individual patient-level data. Models with random intercepts per cohort and fixed effects for age, sex, and years of education were fitted for each outcome of interest in analysis A. Based on the findings from these models, post hoc evaluation of the 2-step meta-analysis in sex subgroups was performed.

TABLE 1. Demographic and clinical characteristics of cohorts

	PD								HC			
Cohort	n	Age, mean (SD)	Male, %	Years of education, mean (SD)	UPDRSs-III, median (IQR)	IPMDS UPDRS-III, median (IQR)	PD disease duration (years), median (IQR)	PD Symptom duration (years), median (IQR)	n	Age, median (SD)	Male, %	Years of education, mean (SD)
California (San Diego)	102	68.1 (7.70)	68%	16.5 (2.58)	21 (15, 31)	—	—	4.0 (2.0, 8.0)	47	66.9 (8.91)	47%	16.2 (2.55)
Pennsylvania (Udall)	247	69.8 (7.7)	67%	16.0 (2.63)	22 (14, 29)	—	—	7.0 (4.0, 10.3)	—	—	—	—
Pennsylvania (K23)	168	63.5 (10.43)	69%	16.2 (2.99)	19 (14, 26)	—	4.0 (2.0, 8.3)	—	—	—	—	—
Arizona (Azsand)	130	73.4 (8.84)	62%	15.5 (2.82)	21 (12, 34)	—	6.4 (3.7, 10.5)	8.0 (4.0, 12.0)	681	78.0 (8.44)	36%	15.2 (2.67)
Arizona (Phoenix)	109	67.3 (6.82)	60%	14.9 (2.75)	—	—	8.0 (5.0, 10.0)	10.0 (7.0, 12.0)	—	—	—	—
Illinois (Chicago)	76	72.7 (6.02)	78%	15.3 (3.16)	28 (23, 35)	—	8.5 (6.8, 12.0)	—	24	71.8 (6.07)	58%	16.4 (2.75)
Washington (Seattle + +Portland)	528	66.5 (9.11)	65%	15.9 (2.50)	—	—	5.0 (2.0, 9.0)	7.0 (4.0, 12.0)	—	—	—	—
Pennsylvania (Hershey)	70	61.3 (8.28)	59%	15.0 (2.65)	—	19 (11, 28)	2.0 (1.0, 6.32	4.3 (2.3, 8.5)	73	59.2 (7.42)	49%	15.6 (2.66)
Canada (Toronto)	138	71.0 (5.43)	67%	15.9 (2.47)	—	26 (20, 33)	4.0 (2.0, 8.0)	6.0 (3.0, 9.8)	—	—	—	—
New Zealand (Christchurch)	160	66.4 (8.24)	66%	12.9 (2.82)	—	31 (23, 42)	2.0 (0.0, 5.0)	4.0 (2.0, 7.0)	63	67.4 (8.56)	67%	13.5 (2.67)
Australia (Sydney)	319	67.6 (9.38)	68%	13.7 (3.45)	—	28 (17, 41)	4.0 (1.5, 9.1)	—	—	—	—	—
Taiwan (Taipei)	70	60.6 (6.01)	63%	11.7 (4.44)	—	—	—	3.0 (2.0, 5.8)	100	65.3 (10.59)	38%	12.3 (3.4)
Netherlands (CARPA)	120	66.0 (10.53)	53%	11.6 (2.47)	15 (11, 22)	—	—	1.3 (1, 2.0)	70	63.7 (7.3)	53%	12.3 (2.44)
Netherlands (VUmc)	248	64.7 (10.2)	61%	13.0(3.26)	25 (18, 35)	—	—	3.0 (2.0, 7.0)	—	—	—	—
Germany (DeNoPa)	159	65.3 (9.7)	66%	9.3 (1.73)	18 (11, 26)	—	De novo	1.3 (0.8, 2.0)	107	64.7 (6.9)	60%	9.6 (1.78)
Spain (Pamplona)	50	69.2 (5.15)	70%	9.8 (3.83)	13 (10, 21)	—	—	12.0 (10.3, 14.0)	62	70.5 (5.05)	48%	9.4 (3.83)
Spain (Barcelona)	99	63.8 (10.94)	59%	10.4 (5.44)	13 (9, 23)	—	—	7.0 (3.5, 12.0)	42	64.9 (11.07)	50%	11.2 (4.53)
Italy (Pisa)	59	73.2 (8.2)	63%	6.6 (3.57)	—	—	3.1 (1.1, 5.0)	—	19	69.6 (7.52)	37%	6.8 (3.42)
Italy (Salerno)	67	59.5 (8.42)	64%	11.5 (3.95)	—	—	1.0 (1.0, 2.0)	—	—	—	—	—
Italy (Venice)	182	63.4 (9.56)	67%	11.0 (4.67)	—	—	8.2 (5.0, 12.9)	—	45	57.3 (9.85)	38%	12.4 (4.2)

TABLE 2. Neuropsychological subtest details

Test outcome	Domain	PD only (number of cohorts)	PD and HC			Original or primary reference(s)[a]
		PD, n	PD, n	HC, n	Cohorts, n	
Trail-Making Test A	AW	2109 (11)	637	224	5	Reitan RM. Validity of the Trail Making test as an indicator of organic brain damage. Percept. Mot Skills. 1958;8:271-276.
Trail-Making Test B	AW or E	2100 (11)	637	224	5	
Stroop Word	AW	1285 (10)	731	1038	7	Stroop JR. Studies of interference in serial verbal reactions. J Exp Psychol. 1935;18 (6):643-662.
Stroop Color	AW	1285 (10)	731	1038	7	
Stroop Interference	AW or E	1352 (11)	731	1038	7	
Digit Span total	AW	1689 (12)	707	409	7	Wechsler D. Wechsler Adult Intelligence Scale – Third Edition: Administration and Scoring Manual. San Antonio, TX: The Psychological Corporation; 1997.Wechsler D. Wechsler Adult Intelligence Scale – Revised. New York: Psychological Association; 1981.
Digit Span Forward	AW	648 (4)	338	196	3	
Digit Span Backward	AW	508 (4)	—	—	—	
WAIS Digit-Symbol Coding	AW	1023 (6)	—	—	—	
WAIS Matrix Reasoning	E	339 (3)	—	—	—	
WAIS Vocabulary	L	278 (3)	—	—	—	
WAIS Similarities	L	607 (5)	607	346	5	
WMS letter-number	AW	1059 (5)	365	236	4	Wechsler D. Wechsler Memory Scale. Third edition manual. San Antonio, TX: The Psychological Cooperation; 1997a.
WMS logical memory I	M	676 (5)	248	171	3	
WMS logical memory II	M	806 (6)	248	171	3	
HVLT immediate recall	M	1370 (6)	—	—	—	Brandt J.. The Hopkins Verbal Learning Test: Development of a new verbal memory test with six equivalent forms. Clin Neuropsychol 1991;5:125-142.
HVLT delayed recall	M	1370 (6)	—	—	—	
HVLT recognition	M	594 (4)	—	—	—	
HVLT delayed recall %	M	663 (3)	—	—	—	
RAVLT immediate recall	M	317 (3)	—	—	—	Rey A. L'Examen Clinique en Psychologie. Paris: Presses Universitaires de France; 1964.
RAVLT delayed recall	M	317 (3)	—	—	—	
CVLT immediate recall	M	400 (3)	—	—	—	Delis DC, Kramer JH, Kaplan E, Ober BA. Manual for the California Verbal Learning Test, (CVLT-II). San Antonio, TX: The Psychological Corporation; 2000.
CVLT short-delayed recall	M	400 (3)	—	—	—	
CVLT long-delayed recall	M	400 (3)	—	—	—	
Animal fluency	E	2908 (17)	1151	1247	10	Henley NM. A psychological study of the semantics of animal terms. J Verbal Learning Verbal Behav 1969;8(2):176-184.
Controlled oral word association	E	2507 (15)	1098	1152	9	Benton AL, Hamser K deS. Multilingual Aphasia Examination (3rd ed.). Iowa City, IA: AJA; 1989.
MWCST perseverative error	E	401 (4)	292	217	3	Nelson HE. A modified card sorting test sensitive to frontal lobe defects. Cortex 1976;12:313-324.
MWCST nonperseverative error	E	292 (3)	292	217	3	
BNT	L	913 (7)	289	185	3	Kaplan, Goodglass H, Weintraub S. Boston Naming Test. Philadelphia: Lea & Febiger; 1983.
JOLO	V	1399 (8)	419	866	4	Benton AL, Sivan AB, Hamsher KS, Varney NH, Spreen O. Contributions to neuropsychological assessment. A clinical manual. New York: Oxford University Press; 1994.

a) Listed are key references for the included tests, but the specific test versions, translations, and normative reference materials utilized varied across studies.
Abbreviations: CVLT = California Verbal Learning Test; HVLT = Hopkins Verbal Learning Test; MWCST = Modified Wisconsin Card Sorting Test; RAVLT = Rey Auditory Verbal Learning Test; WAIS = Wechsler Adult Intelligence Scale; WMS = Wechsler Memory Scale; BNT = Boston Naming Test; JOLO = Judgement of Line Orientation; AW = attention/working memory, E = executive function, M = memory, L = language, V = visuospatial function

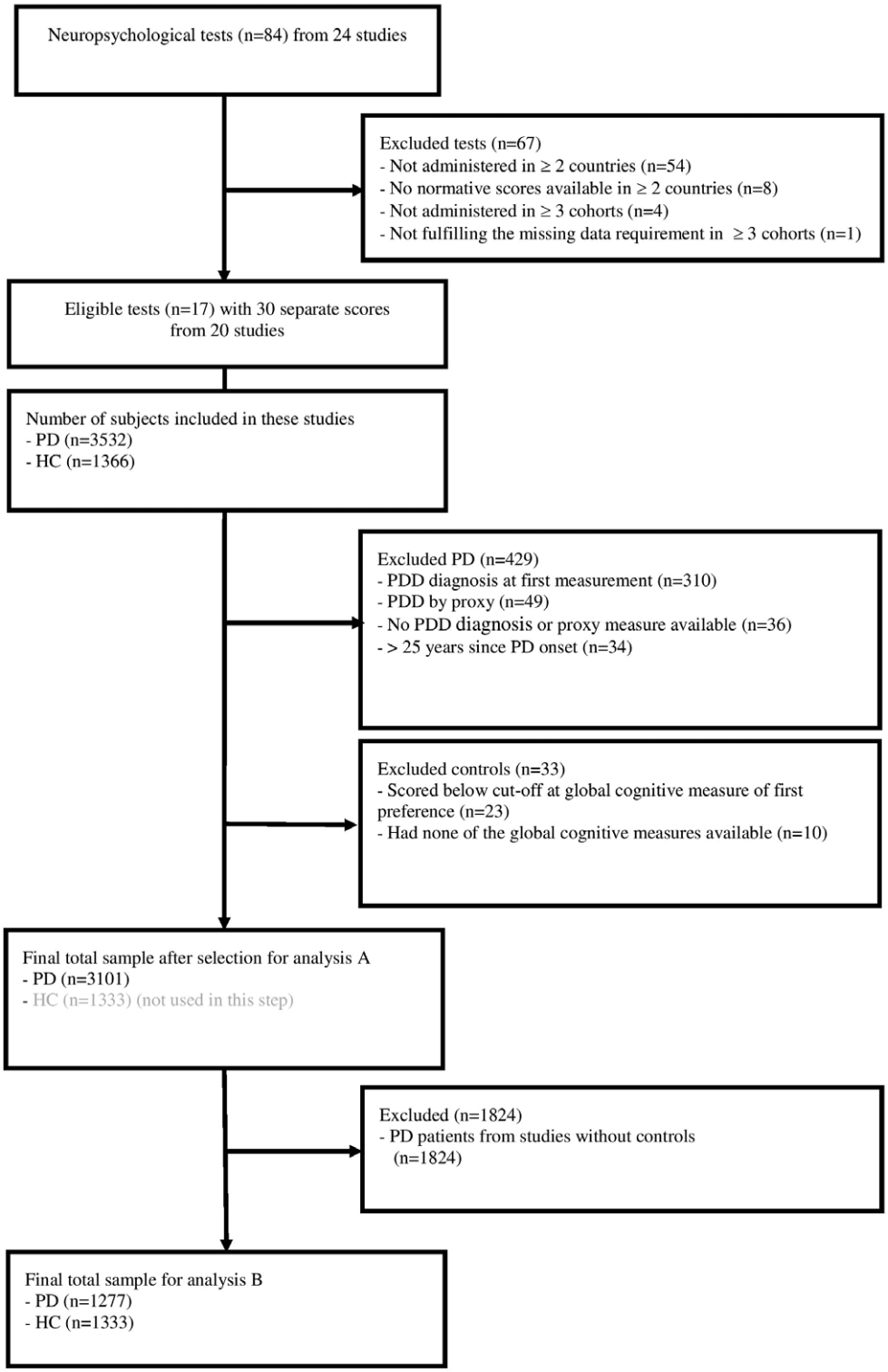

FIG. 1. Flowchart.

Analysis B: Neuropsychological Performance in PD Versus HCs Using Published Norms

This analysis compared performance on normative scores in PD patients relative to the performance of the HCs, with the latter enrolled at the same sites as the respective PD patient samples. The tests available in both PD patients and HCs were a subset of those used for analysis A (Table 2). As in analysis A, a 2-step approach was used. Pooled estimates of neuropsychological performance in PD and HCs were obtained within each of the

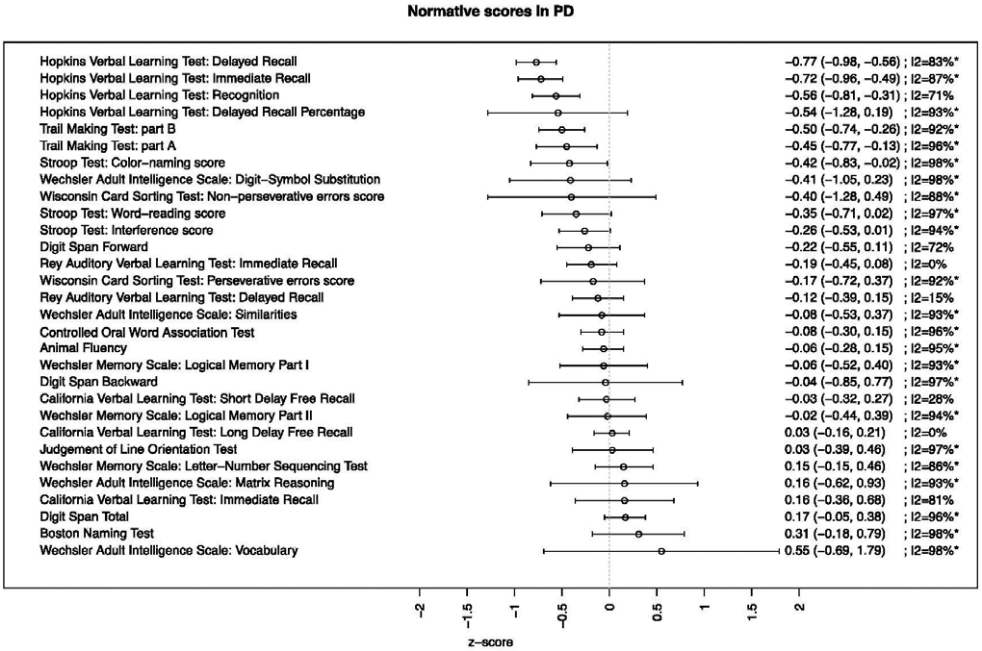

FIG. 2. Pooled PD results per subtest in all participants (analysis A). A star behind the numerical estimates indicates that the lower bound of the heterogeneity index was above 50% (ie, at least half the variability in the final estimates was a result of between-study variability).

original studies and entered into both fixed- and random-effects meta-analysis models using the Hartung Knapp variance estimator to construct the confidence intervals accounting for between-study variability.

In addition, individual patient data were used to correct the results of analysis B for possible differences in age, sex, and education between the PD and HC cohorts. Models with random intercepts and patient-versus-control effects per cohort and fixed effects for age, sex, and years of education were fitted for each outcome of interest in analysis B.

Analysis C: Neuropsychological Performance in HCs Using Published Norms

These methods were exactly similar to the 2-step meta-analyses used for analyses A, but applied to HCs instead of PD participants.

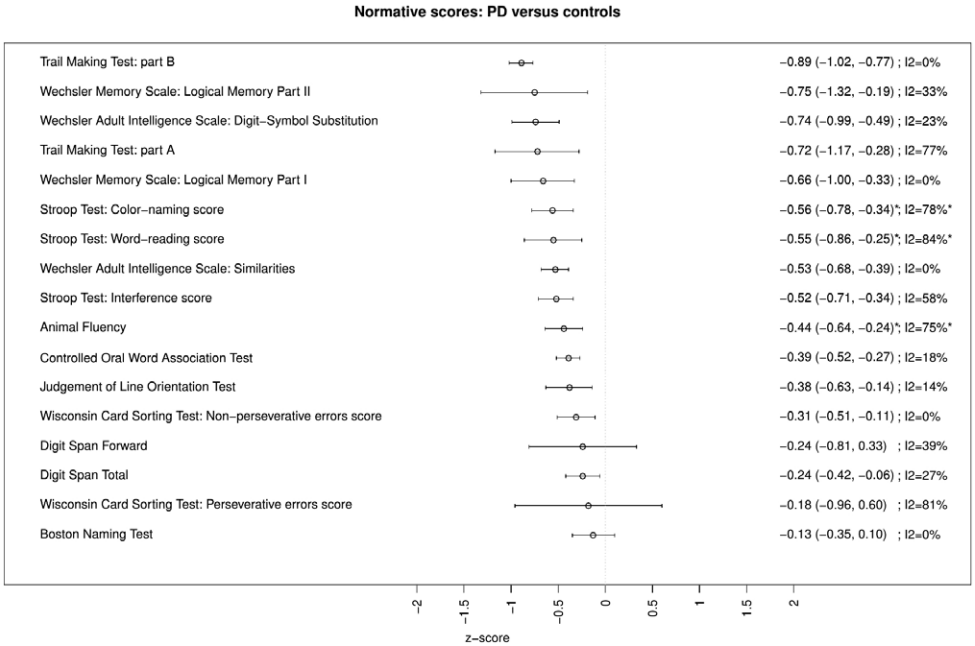

FIG. 3. Pooled PD versus HC results per subtest (analysis B). A star behind the numerical estimates indicates that the lower bound of the heterogeneity index was above 50%.

Software

All analyses were performed using the R statistical software.[32] SPSS 22.0[33] was used for data management.

Results

Analysis A (Neuropsychological Performance in PD Using Published Norms)

The total available number of PD patients was 3101 (Fig. 1). Figure 2 shows overall estimates from the random effects models for each of the tests. In total, 77% of the comparisons showed substantial signs of between-study variability, and the median I^2 over all comparisons in analysis A was 93%. This high degree of between-study variability precluded strong conclusions based on the overall estimates. Nonetheless, despite the observed heterogeneity, the 95% CIs still excluded zero for 6 test scores: 3 verbal memory performance indices of the Hopkins Verbal Learning Test (HVLT), both indices of the Trail-Making Test (TMT), and Stroop Color naming.

Supplementary Figure 1 shows forest plots, including overall random effects model estimates (fixed-effect estimates were omitted because of the observed heterogeneity), measures of heterogeneity, and study-specific means, standard errors, numbers of observations, and analyses weights for each individual cognitive test. To illustrate the observed heterogeneity, the Animal Fluency forest plot shows that many study-specific results deviate from zero on both sides. Prediction of results for a future study would be unreliable because most of the variability depends on unaccounted study characteristics.

Despite using standardized scores, residual effects of age, sex, and education were found on many of the test scores (Supplementary Table 2). Increasing age (14 test scores), male sex (19 test scores), and fewer years of education (25 test scores) were associated with worse performance. Exceptions were relatively better performance for men on Judgement of Line Orientation (JOLO) and Boston Naming Task and for increasing age on WAIS vocabulary. The residual effects of age were relatively small (ie, up to 0.31 SD change for a 10-year age difference), whereas some of the sex and education effects were quite large (eg, women scoring 0.7-0.8 SD lower on the HVLT- and RAVLT immediate recall and each year of education adding 0.15 SD to WAIS similarities performance). In general, these differences between age, sex, and education effects can be explained by the current state of normative data for which correction for age is always applied, but only a minority correct for all 3 characteristics.

Results for the post hoc meta-analyses in sex subgroups are shown in Supplementary Figure 2 (women) and Supplementary Figure 3 (men). Heterogeneity was reduced but still problematic. Female-specific results only showed significant deviation from the reference population mean for TMT A and B. Male-specific results showed deviations for 10 test scores: 5 verbal memory performance indices (3 from the HVLT and 2 from the Rey Auditory Verbal Learning Test [RAVLT]), TMT A and B, and Stroop Word, Color, and Interference.

Analysis B (Neuropsychological Performance in PD Versus HCs Using Published Norms)

A total of 1226 HCs were available (Fig. 1). Figure 3 shows the estimates for the PD-versus-HC normative score comparisons. On average, PD patients differed from HCs on all but 3 of the evaluated measures (Digit Span Forward, Modified Wisconsin Card Sorting Test perseverative errors, and Boston Naming Test). The degree of between-study variability was smaller than in analysis A, with 18% of I^2 CIs excluding 50% on the lower side. This is illustrated in Supplementary Figure 4 by the forest plots for analysis B.

The test scores with 95% CIs excluding zero included test scores in all 5 cognitive domains specified in the PD-MCI criteria: language (WAIS Similarities); attention and working memory (WAIS Digit-Symbol Substitution [or Coding], TMT A, 2 Stroop scores [Word and Color], and Digit Span total); executive (letter fluency [COWAT], animal fluency, Stroop Interference, TMT B, and Modified Wisconsin Card Sorting Test nonperseverative errors); memory (both WMS logical memory indices); and visuospatial (JOLO).

Correcting the analyses for overall effects of age, sex, and education yielded the same results, although quantitatively slightly different (most notably, the estimate for Digit Span Forward was more precise), as shown in Supplementary Figure 5. The effects of age, sex, and education are summarized in Supplementary Table 3 and show worse performance for increasing age on animal fluency, TMT B, Stroop Interference, and JOLO; male sex on animal fluency, COWAT, Stroop Word, Color, and Interference, WAIS-III Digit-Symbol Substitution (or Coding), and WMS logical memory part I/II; female sex on JOLO; and lower education on all test scores. As for analysis A, the age effects were relatively small, whereas some sex and education effects were substantial.

Analysis C (Neuropsychological Performance in HCs Using Published Norms)

To put the performance of PD patients relative to HCs in perspective, we examined HC normative performance separately. Examining the 14 test scores that PD patients demonstrated impairment relative to HCs, the HCs performed above average (ie, the 95% CI excluded zero) on 50% of the tests (7 of 14) and borderline (95% CI bordered on zero) on 14% (2 of 14); see Supplementary Figure 6.

Discussion

In our analyses of a large, multisite, international database of nondemented PD patients, we identified a number of neuropsychological tests across 5 cognitive domains that consistently showed cognitive deficits in nondemented PD patients compared with HCs. The low scores demonstrated by PD patients in our cohorts across multiple cognitive domains are consistent with recent research demonstrating that nondemented PD patients show impairment in memory as well as in the more commonly reported domains of attention and executive function.[34] Our results also showed some impairment in language and visuospatial tests.

An unanticipated result was the high degree of between-study variability in cognitive performance when using normative scores in PD patients without inclusion of site-specific HC data. This variability was despite normative scores accounting for age; however, some of the normative databases used at individual sites are not the most current available, normative scores less commonly corrected for sex and education effects, and some of the PD cohorts included in this study had high mean years of formal education. Some variability was expected because of differences in populations sampled, study conduct, test versions used, and normative reference populations, but the between-study variability was larger than the within-study variability for the majority of test scores examined. Unfortunately, the limited number of studies using each test prevented further post hoc search for possible useful subsets of studies showing more homogeneous results (eg, incidence cohorts only, use of same language). Only the sex subgroups provided sufficient data to do subgroup analyses, and interpretation of observed sex group differences was almost equally hampered by heterogeneity. Furthermore, power differences because of female sample sizes being only approximately 50% as large as male sample sizes further precluded direct comparison. Overall, our results show that normative test scores for the tests currently pooled together may not be comparable across different studies without taking other sources of variability into account, such as unaccounted-for demographic or clinical effects on PD performance across sites, and testing language, version (eg, WAIS-R versus WAIS-III) and normative population. The possibly large influence of test-related characteristics especially highlights the need to further unify existing guidelines on allowed test procedures for operationalization of PD-MCI and PDD criteria.

We attribute the more homogeneous results of the comparisons involving HCs to matching-within-study conditions, including language, test version, test procedures, source population, and educational background. However, although results were more consistent, unrepresentativeness of HCs across studies may have biased those results, and our finding of overall above-average cognitive performance in HCs may reflect this (ie, enrollment of "supercontrols"). Therefore, although the pooled measure of the difference between PD and site-specific HCs is a relatively good summary of the available data, especially when compared with the pooled results based only on normative test scores, the HC data might not be representative of the local general population. Although inclusion of an HC group is a time-consuming and costly task and as noted HC participants need to be representative of population of interest to obtain valid results, inclusion of site-specific HCs can add value to a study in terms of power and comparability with other studies by minimization of nuisance variability related to study conditions. Furthermore, creating extensive normative data sets that account for age, education, and sex will enhance the quality of neuropsychological assessment and can add to the comparability of (international) studies.

Study limitations include post hoc study design; the limited number of neuropsychological tests that met inclusion criteria; the necessity to pool different versions of tests; inclusion of only tests for which standardized scores were available; variable correction for sex, age, and education across the different tests when generating standardized scores; the limited overlap between the tests administered in different studies; and the limited possibility to account for between-study heterogeneity. Coadministration of the proposed tests over time across multiple cohorts that include HCs would enable evaluation of their unique contributions, which ideally would be high to demonstrate that they indeed assess different cognitive abilities.

Because of these limitations, one original study goal, to assemble a preliminary battery of detailed neuropsychological tests administered together to detect cognitive impairment in PD, could not be met. The degree of variability in tests and test versions used was larger than anticipated, and the resulting low degree of overlap between the different studies limited joint evaluation of the tests. Moreover, the incongruence between findings using only published norms versus findings incorporating site-specific HC data and the above-average performance of the HCs in general hindered easy interpretation of the results. However, it is important to note that the lack of definitive results in part reflects the diversity in testing procedures across the world. As such, our findings highlight the need to validate a PD-specific cognitive battery for use internationally.

In this large, multisite, international study of nondemented PD patients, a group of neuropsychological tests was identified that captured impaired cognitive performance in PD patients when compared with directly-assessed site-specific HCs. These tests cover all 5 primary cognitive domains recommended for assessment in nondemented PD patients.[14] However, because of the above-average performance in HCs, these

findings might not reflect absolute deficits from the general population. Moreover, cognitive performance in PD, as measured based on published norms (without a reference group), revealed large between-study variability to an extent that undermines adequate characterization of the data by pooled estimates. Thus, based on these data, it is not possible to recommend with confidence a test battery that would be sensitive to detect mild cognitive deficits in PD patients across multiple international sites. Future research should determine the circumstances under which comparability in test performance across cohorts is expected and warranted and to identify a cognitive battery comprising such tests that are sensitive to cognitive decline, responsive to therapeutic interventions, and superior to individual cognitive tests.

References

1. Aarsland D, Andersen K, Larsen J, Lolk A, Kragh-Sørensen P. Prevalence and characteristics of dementia in Parkinson disease: an 8-year prospective study. Arch Neurol 2003;60:387-392.
2. Litvan I, Aarsland D, Adler C, et al. MDS Task Force on mild cognitive impairment in Parkinson's disease: critical review of PD-MCI. Mov Disord 2011;26:1814-1824.
3. Hely M, Reid W, Adena M, Halliday G, Morris J. The Sydney multicenter study of Parkinson's disease: the inevitability of dementia at 20 years. Mov Disord 2008;23:837-844.
4. Emre M, Aarsland D, Albanese A, et al. Rivastigmine for dementia associated with Parkinson's disease. N Engl J Med 2004;351: 2509-2518.
5. Dubois B, Tollefson G, Katzenschlager R, et al. Donepezil in Parkinson's disease dementia: a randomized, double-blind efficacy and safety study. Mov Disord 2012;27:1230-1238.
6. Emre M, Tsolaki M, Bonuccelli U, et al. Memantine for patients with Parkinson's disease dementia or dementia with Lewy bodies: a randomized, double-blind, placebo-controlled trial. Lancet Neurol 2010;9:969-977.
7. Aarsland D, Ballard C, Walker Z, et al. Memantine in patients with Parkinson's disease dementia or dementia with Lewy bodies: a double-blind, placebo-controlled, multicentre trial. Lancet Neurol 2009;8:613-618.
8. Pigott K, Rick J, Xie S, et al. Longitudinal study of normal cognition in Parkinson disease. Neurology 2015;85:1276-1282.
9. Hoogland J, Boel JA, de Bie RMA, et al. Mild cognitive impairment as a risk factor for Parkinson's disease dementia. Mov Disord 2017; 32:1056-1065.
10. Pedersen KF, Larsen JP, Tysnes OB, Alves G. Natural course of mild cognitive impairment in Parkinson disease: A 5-year population-based study. Neurology 2017;88:767-774.
11. Pedersen K, Larsen J, Tysnes O-B, Alves G. Prognosis of mild cognitive impairment in early Parkinson disease: the Norwegian ParkWest Study. JAMA Neurology 2013;70:580-586.
12. Wood K-L, Myall DJ, Livingston L, et al. Different PD-MCI criteria and risk of dementia in Parkinson's disease: 4-year longitudinal study. NPJ Parkinsons Dis 2016;2:15027.
13. Goldman JG, Weintraub D. Advances in the treatment of cognitive impairment in Parkinson's disease. Mov Disord 2015;30: 1471-1489.
14. Litvan I, Goldman J, Troster A, et al. Diagnostic criteria for mild cognitive impairment in Parkinson's disease: Movement Disorder Society Task Force Guidelines. Mov Disord 2012;27:349-356.
15. Watson G, Cholerton B, Gross R, et al. Neuropsychologic assessment in collaborative Parkinson's disease research: a proposal from the National Institute of Neurological Disorders and Stroke Morris K. Udally Centers of Excellence for Parkinson's Disease Research at the University of Pennsylvania and the University of Washington. Alzheimers Dement 2013;9:609-614.
16. Grinnon ST, Miller K, Marler JR, et al. National Institute of Neurological Disorders and Stroke Common Data Element Project - approach and methods. Clin Trials 2012;9:322-329.
17. Geurtsen GJ, Hoogland J, Goldman JG, et al. Parkinson's disease mild cognitive impairment: application and validation of the criteria. J Parkinsons Dis 2014;4:131-137.
18. Skorvanek M, Goldman J, Jahanshahi M, et al. Rating scales for cognition in Parkinson's disease: critique and recommendations. Mov Disord 2017;33(2):208-218.
19. Dalrymple-Alford J, MacAskill M, Nakas C, et al. The MoCA: well-suited screen for cognitive impairment in Parkinson disease. Neurology 2010;75:1717-1725.
20. Nasreddine Z, Phillips N, Bédirian V, et al. The Montreal Cognitive Assessment, MoCA: a brief screening tool for mild cognitive impairment. J Am Geriatr Soc 2005;53:695-699.
21. Brown GG, Rahill AA, Gorell JM, et al. Validity of the Dementia Rating Scale in assessing cognitive function in Parkinson's disease. J Geriatr Psychiatry Neurol 1999;12:180-188.
22. Matteau E, Dupre N, Langlois M, Provencher P, Simard M. Clinical validity of the Mattis Dementia Rating Scale-2 in Parkinson disease with MCI and dementia. J Geriatr Psychiatry Neurol 2012;25: 100-106.
23. Folstein M, Folstein S, McHugh P. "Mini-Mental State": A practical method for grading the cognitive state of patients for the clinician. J Psychiatr Res 1975;12:189-198.
24. van Buuren S, Groothuis-Oudshoorn K. Mice: multivariate imputation by chained equations in R. J Stat Softw 2011;45:1-67.
25. van Buuren S. *Flexible Imputation of Missing Data*. Boca Raton, FL: Chapman and Hall/CRC; 2012.
26. Rubin D. Inference and missing data. Biometrika 1976;63: 581-592.
27. Burgess S, White IR, Resche-Rigon M, Wood AM. Combining multiple imputation and meta-analysis with individual participant data. Stat Med 2013;32:4499-4514.
28. Marshall A, Altman DG, Holder RL, Royston P. Combining estimates of interest in prognostic modelling studies after multiple imputation: current practice and guidelines. BMC Med Res Methodol 2009;9:57.
29. Riley RD, Lambert PC, Abo-Zaid G. Meta-analysis of individual participant data: rationale, conduct, and reporting. BMJ 2010;340: c221.
30. Schwarzer G. Meta: an R package for meta-analysis. R News 2007; 7:40-45.
31. Hartung J, Knapp G. On tests of the overall treatment effect in meta-analysis with normally distributed responses. Stat Med 2001; 20:1771-1782.
32. *R: A language and environment for statistical computing [computer program]*. Vienna, Austria: R Foundation for Statistical Computing; 2014.
33. *IBM SPSS Statistics, for Windows, version 22.0 [computer program]*. Armonk, NY: IBM Corp.; 2013.
34. Aarsland D, Bronnick K, Williams-Gray C, et al. Mild cognitive impairment in Parkinson's disease: A multicenter pooled analysis. Neurology 2010;75:1062-1069.

Supporting Data

Additional Supporting Information may be found in the online version of this article at the publisher's web-site.