

©The Author(s) 2026. Published by the Radiological Society of North America under a CC BY 4.0 license.

10.1148/radiol.251394

**Table S1.** Study cohort details.

<b>Data</b>	<b>PSP-RS Group (n=563)</b>	<b>PSP variant Group (n=104)</b>	<b>MSA Group (n=93)</b>	<b>CBS Group (n=98)</b>	<b>PD Group (n=643)</b>	<b>Control Participant Group (N=610)</b>
<i>Italian Training Cohort (n=459)</i>						
University of Catanzaro, Italy (n=459)	93	43	30	/	208	85
<i>International external Test Cohort (n=1609)</i>						
Gosuranemab Trial (n=157)	153	4	/	/	/	/
Davunetide (AL-108-231) Trial (n=144)	144	/	/	/	/	/
Tideglusib Trial (n=9)	4	5	/	/	/	/
DescribePSP German cohort (n=32)	27	5	/	/	/	/
4RTNI (n=110)	62	9	/	39	/	/
PROSPECT-UK cohort (n=70)	23	14	17	16	/	/
University of Athens, Greece (n=277)	46	24	40	43	43	81
PPMI (n=493)	/	/	/	/	366	127
ADNI (n=202)	/	/	/	/	/	202
OASIS (n=115)	/	/	/	/	/	115
<i>Pathologically proven cohort (n=43)</i>						
4RTNI (n=4)	4	/	/	/	/	/
PPMI (n=22)	/	/	1	/	21	/
University of Turku, Finland (n=17)	7	/	5	/	5	/

Abbreviations: PSP-RS = progressive supranuclear palsy - Richardson's Syndrome; MSA = Multiple system atrophy; CBS = Cortico-basal syndrome; PD = Parkinson's disease. 4RTNI = 4R-Tau Neuroimaging Initiative; PPMI = Parkinson's Progression Markers Initiative; ADNI = Alzheimer's Disease Neuroimaging Initiative; OASIS = Open Access Series of Imaging Studies. Overall, the international external test cohort included 459 PSP-RS, 61 PSP variants, 57 MSA, 98 CBS, 409 PD and 525 control participants.

**Table S2.** Classification performances of midbrain measures in distinguishing participants with PSP from those with non-PSP parkinsonism. Linear measurements of midbrain (A or B) were increased in participants with PSP and/or decreased in those with non-PSP parkinsonism by 10% or 20% to evaluate the effect of possible measurement errors on classification performances.

<b>1a) PSP vs non-PSP      A or B measures increased in the PSP group</b>			
<b>ERROR 10%</b>	<b>A values increased by 10%</b>	<b>B values not modified</b>	<b>DMPI</b>
AUC	0.90		0.92
ACC	0.84	-	0.87
SENS	0.82		0.85
SPEC	0.85		0.88
	<b>A values not modified</b>	<b>B values increased by 10%</b>	<b>DMPI</b>
AUC		0.88	0.93
ACC	-	0.81	0.88
SENS		0.80	0.87
SPEC		0.81	0.90
<b>ERROR 20%</b>	<b>A increased by 20%</b>	<b>B values not modified</b>	<b>DMPI</b>
AUC	0.80		0.89
ACC	0.74	-	0.83
SENS	0.73		0.81
SPEC	0.74		0.85
	<b>A values not modified</b>	<b>B increased by 20%</b>	<b>DMPI</b>
AUC		0.82	0.91
ACC	-	0.75	0.86
SENS		0.75	0.83
SPEC		0.76	0.88
<b>1b) PSP vs non-PSP      A or B measures decreased in the non-PSP group</b>			
<b>ERROR 10%</b>	<b>A values decreased by 10%</b>	<b>B values not modified</b>	<b>DMPI</b>
AUC	0.89		0.92
ACC	0.82	-	0.87
SENS	0.81		0.85
SPEC	0.84		0.88
	<b>A values not modified</b>	<b>B values decreased by 10%</b>	<b>DMPI</b>
AUC		0.87	0.93
ACC	-	0.80	0.88
SENS		0.80	0.87
SPEC		0.80	0.90

<b>ERROR 20%</b>	<b>A values decreased by 20%</b>	<b>B values not modified</b>	<b>DMPI</b>
AUC	0.74		0.87
ACC	0.68	-	0.82
SENS	0.68		0.80
SPEC	0.67		0.84
	<b>A values not modified</b>	<b>B values decreased by 20%</b>	<b>DMPI</b>
AUC		0.79	0.91
ACC	-	0.72	0.86
SENS		0.72	0.84
SPEC		0.72	0.88
<b>1c) PSP vs non-PSP A or B measures increased in PSP and decreased in non-PSP groups</b>			
<b>ERROR 10%</b>	<b>A values modified by 10% in both PSP and non-PSP</b>	<b>B values not modified</b>	<b>DMPI</b>
AUC	0.77		0.88
ACC	0.71	-	0.83
SENS	0.70		0.81
SPEC	0.71		0.85
	<b>A values not modified</b>	<b>B values modified by 10% in both PSP and non-PSP</b>	<b>DMPI</b>
AUC		0.80	0.91
ACC	-	0.74	0.86
SENS		0.73	0.84
SPEC		0.75	0.88
<b>ERROR 20%</b>	<b>A values modified by 20% in both PSP and non-PSP</b>	<b>B values not modified</b>	<b>DMPI</b>
AUC	0.73		0.75
ACC	0.68	-	0.68
SENS	0.66		0.69
SPEC	0.70		0.68
	<b>A values not modified</b>	<b>B values modified by 20% in both PSP and non-PSP</b>	<b>DMPI</b>
AUC		0.64	0.85
ACC	-	0.60	0.79
SENS		0.62	0.77
SPEC		0.57	0.81

Abbreviations: PSP = progressive supranuclear palsy; DMPI = dual-line midbrain PSP index; AUC = area under the curve; ACC = accuracy; SENS = sensitivity; SPEC = specificity.

Data obtained in the whole cohort (Italian + international cohort). The PSP group included 656 participants (552 PSP-Richardson's syndrome and 104 PSP variants); the non-PSP group included 802 participants (617 PD, 87 MSA, 74 CBS and 24 CBS-AD). The scenarios of small mistakes in the midbrain measures (A and/or B) negatively affecting the classification accuracy were explored in the table. In detail, we hypothesized three scenarios where measurement mistakes could have a deleterious effect on classification accuracy: (i) larger measures in participants with PSP, (ii) smaller measures in participants with non-PSP parkinsonism, and (iii) a combination of both these errors. Thus, actual A or B values were increased in PSP and/or decreased in non-PSP participants by 10% or 20% magnitude. These changes were performed alternatively either for A or B measure, and the DMPI was calculated after measurement adjustment, as specified in each table row. In all cases, the classification performances were assessed by logistic regression analysis including the midbrain measure, age and sex. Performance was evaluated using a stratified 5-fold cross-validation, repeated five times. Accuracy, sensitivity and specificity values were calculated by using a discriminating probability threshold of 50% (probability  $\geq 50\%$  suggestive of PSP and probability  $< 50\%$  suggestive of non-PSP). As the result of averaging two measures, the DMPI was less sensitive to possible measurement mistakes, showing the highest AUC values in all the explored scenarios.

**Table S3.** Demographic, clinical and imaging data of the subcohort used for marker comparison.

Data	Whole Cohort			Subcohort		
	PSP	Non-PSP	p-value	PSP	Non-PSP	p-value
	Group (N=667)	Group (N=834)		Group (N=161)	Group (N=203)	
Sex, (Males/Females)	358 / 309	494 / 340	P=.03 <sup>a</sup>	82 / 79	124 / 79	P=.06 <sup>a</sup>
Age at examination, ys <sup>b</sup>	69.1 ± 7.0	65.6 ± 8.1	P<.001 <sup>c</sup>	68.3 ± 6.7	65.1 ± 8.3	P<.001 <sup>c</sup>
Disease onset, ys <sup>b</sup>	66.9 ± 6.9	62.3 ± 8.3	P<.001 <sup>c</sup>	65.2 ± 6.6	62.0 ± 8.7	P<.001 <sup>c</sup>
Disease duration, ys <sup>b</sup>	3.0 ± 2.6	3.3 ± 2.1	P=.20 <sup>c</sup>	3.2 ± 2.7	3.1 ± 2.8	P=.48 <sup>c</sup>
MDS-UPDRS-III score <sup>b</sup>	38.6 ± 15.9	25.2 ± 12.7	P<.001 <sup>c</sup>	34.0 ± 13.5	24.4 ± 12.4	P<.001 <sup>c</sup>
PSPRS score <sup>b</sup>	36.7 ± 12.6	-	-	36.2 ± 12.3	-	-
		643 PD			155 PD	
Disease subtype	563 PSP-RS	93 MSA	-	135 PSP-RS	23 MSA	-
	104 vPSP	74 CBS		26 vPSP	19 CBS	
		24 CBS-AD			6 CBS-AD	
<i>Imaging data</i>						
DMPI, mm <sup>b</sup>	6.48 ± 1.23	9.43 ± 1.16	P<.001 <sup>d</sup>	6.46 ± 1.10	9.41 ± 1.12	P<.001 <sup>d</sup>
Midbrain line, mm <sup>b</sup>	-	-	-	13.5 ± 1.2	16.2 ± 1.3	P<.001 <sup>d</sup>
Midbrain area, mm <sup>2b</sup>	-	-	-	75.2 ± 20.5	125.7 ± 26.3	P<.001 <sup>d</sup>
M/P area ratio <sup>b</sup>	-	-	-	0.16 ± 0.04	0.25 ± 0.08	P<.001 <sup>d</sup>
MRPI <sup>b</sup>	-	-	-	20.4 ± 7.1	11.3 ± 3.0	P<.001 <sup>d</sup>
MRPI 2.0 <sup>b</sup>	-	-	-	4.72 ± 2.31	1.87 ± 0.88	P<.001 <sup>d</sup>

Abbreviations: PD = Parkinson's disease; MSA= Multiple System Atrophy; PSP-RS = Progressive supranuclear palsy-Richardson's syndrome; vPSP = Progressive supranuclear palsy variants; MDS-UPDRS-III = Movement Disorder Society - Unified Parkinson's Disease Rating Scale- part III (Motor Examination); H-Y = Hoehn and Yahr scale; PSPRS= Progressive Supranuclear Palsy Rating Scale.

The sub-cohort was generated by randomly selecting around 25% of individuals from each participant group within the main cohort, maintaining the original balance between Italian and international cohorts. In the sub-cohort, age at onset and disease duration were available for 110 PSP, and 200 non-PSP participants. MDS-UPDRS-III was available for 62 PSP and 173 non-PSP participants; PSPRS score was available for 117 PSP participants. H-Y score was available for 81 PSP-RS, 42 vPSP, 573 PD and 19 MSA participants.

- Fisher's exact test
- Data are expressed as mean ± standard deviation.
- t-tests, or Kruskal-Wallis test,
- ANCOVA with age and sex as covariates.

**Table S4.** ROC classification performances of the DMPI in distinguishing participants with PSP from those with non-PSP parkinsonism.

Performance Metrics	Italian cohort	International cohort	Whole cohort
Cut-off	8.12 (8.00, 8.20)	7.97 (7.62, 8.12)	8.02 (7.97, 8.17)
AUC	0.97 (0.96, 0.99)	0.94 (0.93, 0.96)	0.95 (0.94, 0.96)
Accuracy	361/374, 96.52 (92.78, 97.33)	967/1084, 89.21 (87.36, 90.87)	1321/1458, 90.60 (88.96, 91.98)
Sensitivity	131/136, 96.32 (92.65, 99.26)	461/520, 88.65 (84.04, 91.54)	593/656, 90.39 (87.80, 93.00)
Specificity	225/238, 94.54 (91.60, 97.48)	506/564, 89.72 (86.52, 93.98)	728/802, 90.77 (88.03, 92.77)
PPV	131/144, 90.97 (86.67, 95.59)	461/519, 88.82 (86.03, 92.87)	593/667, 88.90 (86.06, 91.12)
NPV	225/230, 97.82 (95.73, 99.56)	506/565, 89.56 (86.32, 91.93)	728/791, 92.03 (90.10, 93.21)

Abbreviations: PSP = Progressive supranuclear palsy; DMPI = dual-line midbrain PSP index; AUC = Area under the ROC Curve, PPV = Positive predictive value; NPV = Negative predictive value.

The performances of DMPI (without considering age and sex) were calculated using Receiver Operating Characteristic (ROC) analysis. Optimal cut-offs, defined as the values with the highest sum of sensitivity and specificity (Youden's method), and 95% confidence intervals, were calculated using R pROC software package with bootstrapping (n=2,000 iterations). The table shows numerator/denominator, performance metric and 95% confidence intervals. The Italian cohort included 136 PSP and 238 non-PSP, while the international cohort included 520 PSP and 564 non-PSP participants.

**Table S5.** Linear associations between midbrain measures and age or sex.

Data	PSP	PD	Control participant
	Group	Group	Group
A measure – age	-0.14 (P<.001)	-0.45 (P<.001)	-0.48 (P<.001)
A measure – sex	-0.03 (P=.38)	-0.06 (P<.09)	-0.04 (P<.27)
B measure – age	-0.14 (P<.001)	-0.44 (P<.001)	-0.44 (P<.001)
B measure – sex	-0.13 (P<.001)	-0.13 (P<.001)	-0.16 (P<.001)
DMPI – age	-0.15 (P<.001)	-0.48 (P<.001)	-0.50 (P<.001)
DMPI – sex	-0.09 (P=.02)	-0.10 (P=.003)	-0.12 (P<.001)

Abbreviations: PSP = Progressive supranuclear palsy; PD = Parkinson's disease; DMPI = Dual-line midbrain PSP index (calculated by averaging A and B measures). The table shows associations between midbrain measures and demographic variables in the whole cohort (Italian + international cohort, n=2068). Data are beta values (p values) of linear models with the following structure (midbrain measure ~ age + sex), in each group.

**Table S6.** Effect of age and sex on individual probability score of having PSP rather than non-PSP parkinsonism.

Participant	DMPI	Sex and age	Probability of PSP	Prediction	Probability quartile	Probability gap
Participant 1	6.50	Female 45 yo	97.1%	PSP	4 <sup>th</sup>	7.6%
		Male, 80 yo	89.5%	PSP	4 <sup>th</sup>	
Participant 2	7.25	Female 45 yo	89.7%	PSP	4 <sup>th</sup>	20.7%
		Male, 80 yo	69.0%	PSP	3 <sup>rd</sup>	
Participant 3	8.00	Female 45 yo	69.4%	PSP	3 <sup>rd</sup>	32.6%
		Male, 80 yo	36.8%	Non-PSP	2 <sup>nd</sup>	
Participant 4	8.75	Female 45 yo	37.3%	Non-PSP	2 <sup>nd</sup>	24.1%
		Male, 80 yo	13.2%	Non-PSP	1 <sup>st</sup>	
Participant 5	9.50	Female 45 yo	13.5%	Non-PSP	1 <sup>st</sup>	9.7%
		Male, 80 yo	3.8%	Non-PSP	1 <sup>st</sup>	

Abbreviations: PSP = Progressive supranuclear palsy; DMPI = dual-line midbrain PSP index; yo = years old. The probability of having PSP rather than non-PSP parkinsonism was calculated using the logistic regression model trained on the whole cohort (Italian + international), also available at [https://neuroimagingunicz.github.io/mri\\_calc/](https://neuroimagingunicz.github.io/mri_calc/)

The table shows how age and sex can influence the probability scores, based on different combination of demographic variables. As examples, 5 DMPI values were considered, and two scenarios were hypothesized: the case where the DMPI measure was observed in an old man and the case where it was observed in a young woman. As observed, demographic variables have smaller effect on probability scores in presence of very high or low DMPI values, while they have relevant effect (see probability gap) in presence of intermediate DMPI values, potentially leading to changes of probability quartile or even to changes in the diagnosis prediction (PSP or non-PSP, always considering a discriminating threshold of 50%). Age (40 to 100 yo) and sex (male or female) may lead to change in the diagnosis prediction (PSP or non-PSP) for DMPI values between 7.5 and 8.5. In our cohort of 656 PSP and 802 non-PSP participants, DMPI values within this range were observed in 202/1458 (13.8%) of cases, demonstrating the importance of considering age and sex for estimating the probability of having PSP.

**Table S7.** Classification performances of the logistic regression models based on DMPI, age and sex in distinguishing participants with PSP from those with non-PSP parkinsonism in the early stages of the diseases.

Performance Metrics	PSP (n=107) vs non-PSP (n=153) within 1 year from disease onset	PSP (n=209) vs non-PSP (n=353) within 2 years from disease onset	PSP (n=299) vs non-PSP (n=496) within 3 years from disease onset
Probability threshold	0.50	0.50	0.50
AUC	0.96 (95% CI: 0.95, 0.98)	0.97 (95% CI: 0.95, 0.99)	0.96 (95% CI: 0.95, 0.97)
Accuracy	235/260, 90.38 (85.61, 95.16)	510/562, 90.75 (85.74, 95.77)	726/795, 91.32 (89.30, 93.34)
Sensitivity	96/107, 89.72 (83.07, 96.23)	191/209, 91.39 (84.63, 98.20)	271/299, 90.63 (85.78, 95.46)
Specificity	139/153, 90.85 (83.30, 98.29)	319/353, 90.37 (82.50, 98.25)	455/496, 91.73 (89.04, 94.43)

Abbreviations: PSP = Progressive supranuclear palsy; AUC = area under the curve; DMPI = dual-line midbrain PSP index. The DMPI performance was calculated using LR analysis including age and sex in a subset of the whole cohort (Italian + international) including participants at the early stage of the diseases (within 1, 2 or 3 years from the disease onset). The sample size increases as far as we move from left to right panels, since a broader time interval from disease onset was considered. Classification performances were calculated using a logistic regression model including the DMPI value as predictor, and confounding factors (age and sex). Performance was evaluated using a stratified 5-fold cross-validation, repeated five times. Accuracy, sensitivity and specificity were calculated by using a discriminating probability threshold of 50% (probability  $\geq$  50% suggestive of PSP and probability  $<$  50% suggestive of non-PSP).

**Table S8.** Classification performances of the DMPI in pairwise comparisons across the various participant groups

DMPI Performances						
	PSP-RS (#552) vs Controls (#610)	PSP-RS (#552) vs PD (#617)	PSP-RS (#87) vs MSA (#87)	PSP-RS (#104) vs PSPV (#104)	PSP-RS (#74) vs CBS (#74)	PSP-RS (#24) vs CBS-AD (#24)
AUC	0.97 (0.01)	0.97 (0.01)	0.91 (0.04)	0.60 (0.07)	0.77 (0.12)	0.84 (0.13)
ACC	0.93 (0.01)	0.93 (0.01)	0.87 (0.06)	0.61 (0.07)	0.74 (0.07)	0.82 (0.14)
SENS	0.91 (0.01)	0.92 (0.03)	0.83 (0.08)	0.58 (0.10)	0.72 (0.12)	0.77 (0.19)
SPEC	0.95 (0.02)	0.94 (0.01)	0.91 (0.09)	0.64 (0.08)	0.76 (0.08)	0.87 (0.19)

	vPSP (#104) vs Controls (#104)	vPSP (#104) vs PD (#104)	vPSP (#87) vs MSA (#87)	vPSP (#74) vs CBS (#74)	vPSP (#24) vs CBS-AD (#24)
AUC	0.89 (0.04)	0.89 (0.03)	0.92 (0.04)	0.74 (0.08)	0.75 (0.15)
ACC	0.82 (0.04)	0.82 (0.05)	0.86 (0.05)	0.69 (0.07)	0.73 (0.15)
SENS	0.82 (0.05)	0.84 (0.05)	0.83 (0.08)	0.68 (0.13)	0.74 (0.22)
SPEC	0.82 (0.08)	0.81 (0.08)	0.89 (0.09)	0.70 (0.04)	0.71 (0.19)

	PD (#617) vs Controls (#610)	PD (#87) vs MSA (#87)	PD (#74) vs CBS (#74)	PD (#24) vs CBS-AD (#24)
AUC	0.64 (0.05)	0.58 (0.08)	0.75 (0.07)	0.65 (0.14)
ACC	0.60 (0.04)	0.56 (0.06)	0.70 (0.08)	0.64 (0.12)
SENS	0.62 (0.05)	0.59 (0.11)	0.70 (0.11)	0.64 (0.19)
SPEC	0.58 (0.05)	0.54 (0.10)	0.70 (0.11)	0.64 (0.17)

	MSA (#87) vs Controls (#87)	MSA (#74) vs CBS (#74)	MSA (#24) vs CBS-AD (#24)
AUC	0.67 (0.09)	0.74 (0.08)	0.60 (0.11)
ACC	0.64 (0.08)	0.69 (0.09)	0.58 (0.08)
SENS	0.68 (0.11)	0.61 (0.13)	0.46 (0.18)
SPEC	0.60 (0.13)	0.77 (0.09)	0.70 (0.14)

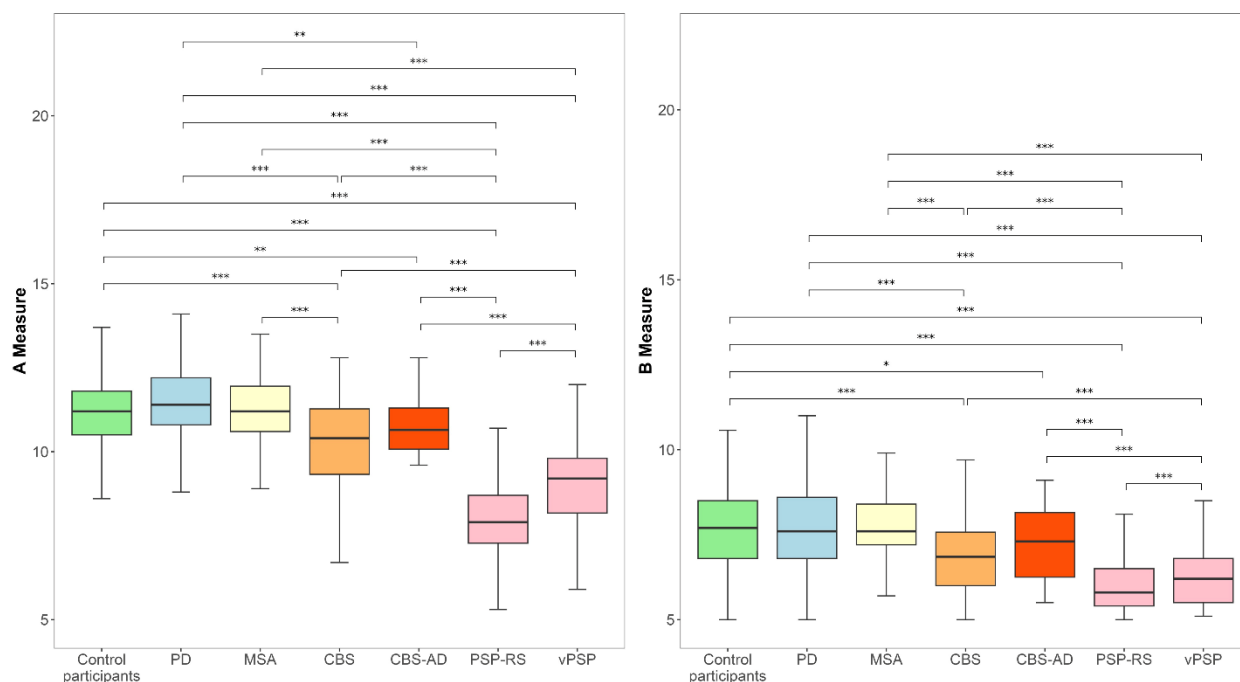
  

	CBS (#74) vs Controls (#74)
AUC	0.77 (0.08)
ACC	0.73 (0.07)
SENS	0.75 (0.08)
SPEC	0.70 (0.13)

Abbreviations: PSP-RS = progressive supranuclear palsy - Richardson's Syndrome; MSA = Multiple system atrophy; CBS = Cortico-basal syndrome; CBS-AD = CBS with cerebrospinal fluid Alzheimer's profile; PD = Parkinson's disease; DMPI = dual-line midbrain PSP index; AUC = area under the curve; ACC = accuracy; SENS = sensitivity; SPEC = specificity. The classification performances

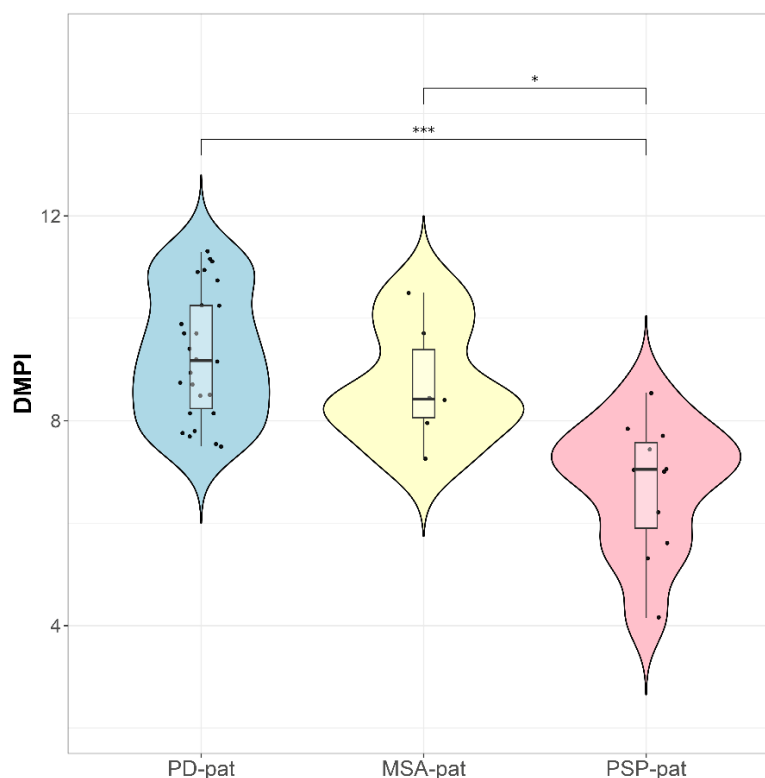
were assessed by logistic regression analysis including the DMPI, age and sex. Performance was evaluated using a stratified 5-fold cross-validation, repeated five times. Accuracy, sensitivity and specificity were calculated by using a discriminating probability threshold of 50% (probability  $\geq 50\%$  suggestive of PSP and probability  $< 50\%$  suggestive of non-PSP). To address class imbalance in the comparisons where one group had a sample size significantly larger than the other one, we performed a cluster-based under-sampling procedure to match the size of the larger group with that of the smaller group, preserving the diversity and distribution of the original data. The under-sampling procedure was repeated ten times and the average classification performances in each comparison are shown in the table.

**Figure S1.**



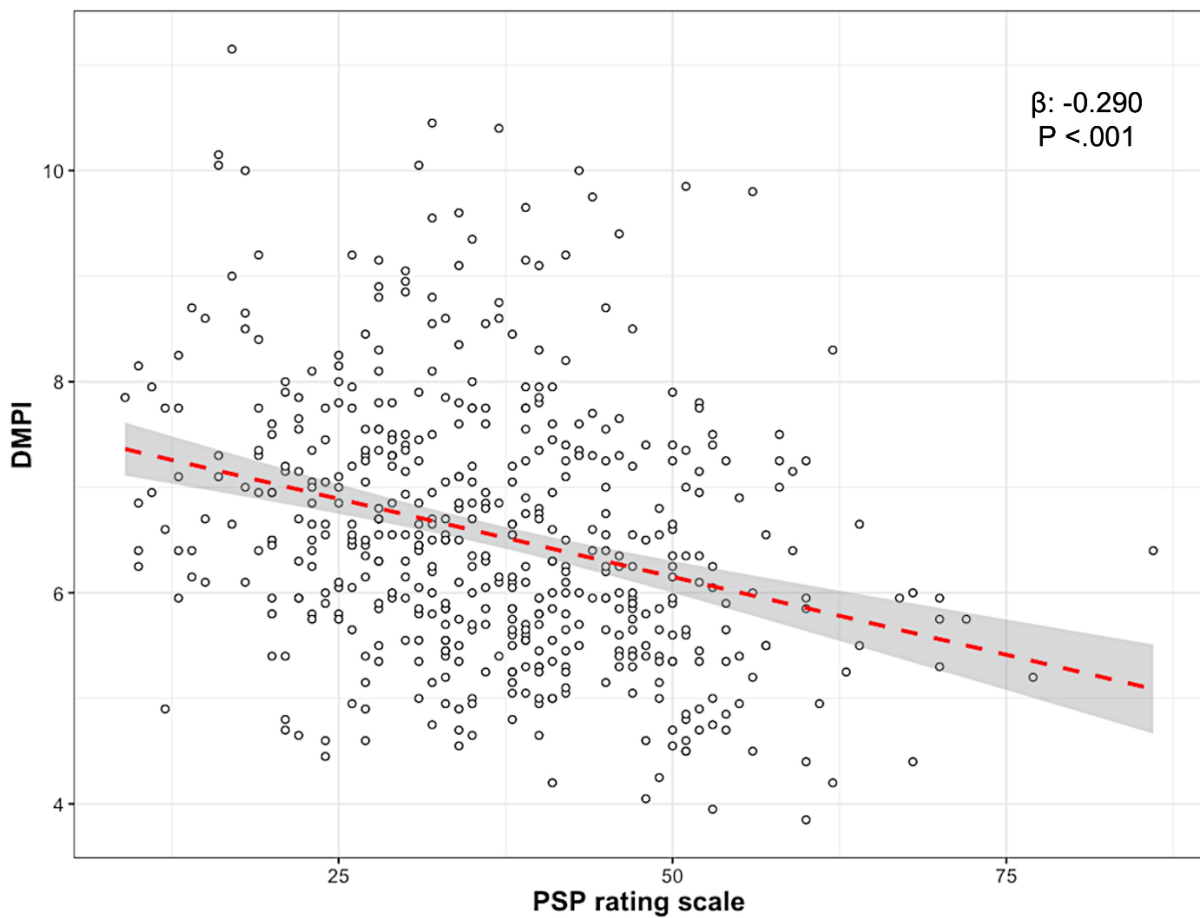
**Figure S1.** The figure shows boxplots of midbrain measures A (left panel) and B (right panel) across different groups (control participants, PD, MSA, CBS, CBS-AD, PSP-RS, vPSP) in the whole cohort (Italian + international cohort, n=2068). Each box represents the interquartile range, with the median shown as a horizontal line. The whiskers extend to the minimum and maximum values, after excluding outliers. For both midbrain measures, ANCOVA with age and sex as covariates, followed by post-hoc comparisons adjusted using Bonferroni correction (for 21 tests) showed significant p-value in PSP-RS vs control participants, PD, MSA, CBS, CBS-AD and vPSP,  $P < .001$ ; vPSP vs control participants, PD, MSA, CBS and CBS-AD,  $P < .001$ ; CBS vs control participants, PD and MSA,  $P < .001$ . No differences were observed among PD, MSA and control participant groups for both measures. Abbreviations: PD = Parkinson's disease; MSA= Multiple System Atrophy; CBS = Cortico-basal syndrome; CBS-AD: CBS with Alzheimer's profile in cerebrospinal fluid; PSP-RS = Progressive supranuclear palsy-Richardson's syndrome; vPSP = Progressive supranuclear palsy variants.

**Figure S2.**



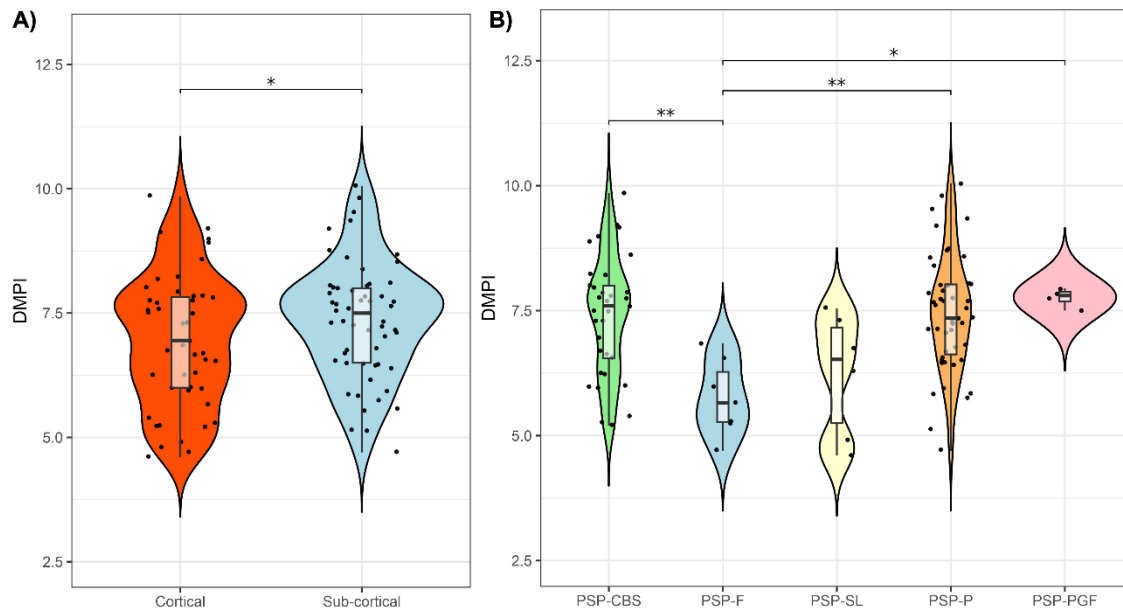
**Figure S2.** Violin plots of DMPI values across different groups (PD, MSA, PSP) in the pathologically confirmed diagnosis cohort (11 PSP, 26 PD, 6 MSA, n=43). Each violin represents the estimated density of the values for each group, highlighting the median (central line in the box plot) and the variability of the data. The overlaid box plots illustrate the IQR and whiskers, while black dots show individual data. ANCOVA was performed with age and sex as covariates, followed by post-hoc comparisons adjusted using Bonferroni correction (for 3 tests), showed significant p-values in the PSP vs PD group ( $P<.001$ ) and PSP vs MSA group ( $P=.04$ ) comparisons. There was no evidence of a difference in DMPI between the PD and MSA groups. \* $P<.05$ , \*\* $P<.01$ , \*\*\* $P<.001$ . When the logistic regression model based on the DMPI value, age and sex trained in the entire study cohort of 656 participants with PSP and 802 participants with non-PSP parkinsonisms (Italian + international cohorts) was applied to this independent cohort of participants with pathologically proven diagnoses, the model AUC was 0.94 (95% CI: 0.86, 1.00) for distinguishing participants with PSP from participants with non-PSP parkinsonisms. With a probability threshold of 50%, the model showed 10/11, 90.91% sensitivity, 27/32, 84.38% specificity and 37/43, 86.05% accuracy. Abbreviations: DMPI = dual-line midbrain PSP index; PD = Parkinson's disease; MSA = multiple system atrophy, PSP = progressive supranuclear palsy; ANCOVA = analysis of covariance; AUC = area under the receiver operating characteristic curve; PD-pat = pathologically confirmed Parkinson's disease; MSA-pat = pathologically confirmed multiple system atrophy; PSP-pat = pathologically confirmed progressive supranuclear palsy.

**Figure S3.**



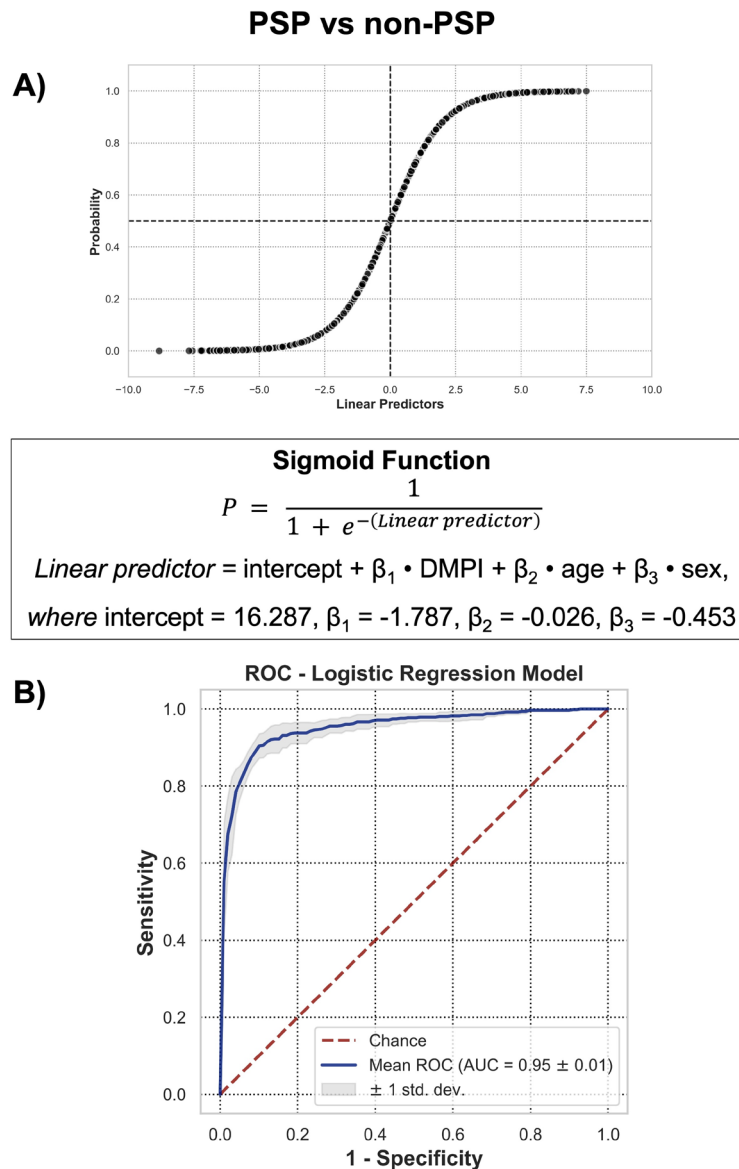
**Figure S3.** The figure shows a significant association between DMPI values and PSP rating scale values in participants with PSP. The PSP group included all participants with PSP from the Italian and international cohorts with available PSP rating scale score (n=513). The  $\beta$  value and p value shown in the figure were derived from linear regression model including age and sex as covariates. Abbreviations: PSP = progressive supranuclear palsy; DMPI = dual-line midbrain PSP index.

**Figure S4.**



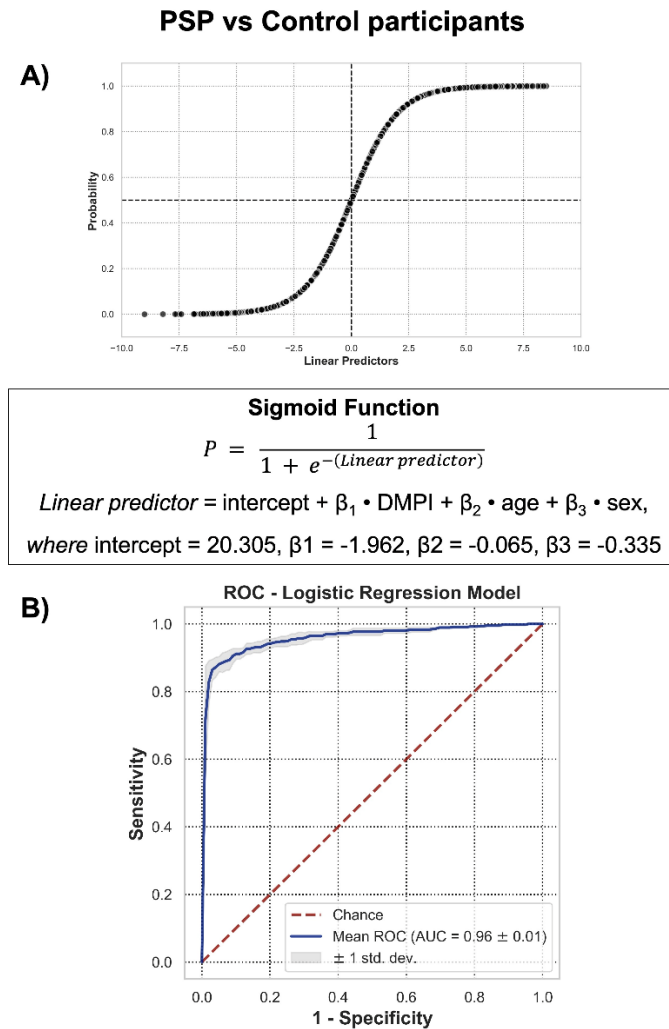
**Figure S4.** The figure shows violin plots of DMPI values across different PSP variants in the whole cohort (n=104). The panel on the left (A) shows data of “cortical” and “subcortical” variant groups, where the former included PSP-CBS, PSP-F and PSP-SL, and the latter included PSP-P and PSP-PGF. The panel on the right (B) shows data of all PSP subtypes in our cohort. Each violin represents the estimated density of the values for each group, highlighting the median (central line in the box plot) and the variability of the data. The overlaid box plots illustrate the interquartile range and whiskers, while black dots show individual data. In the panel A, ANCOVA (covariates: age, sex) showed  $P=.03$ . In the panel B, ANCOVA (covariates: age, sex) followed by post-hoc comparisons adjusted using Bonferroni correction (for 10 tests) showed lower values in PSP-F than in PSP-CBS ( $P=.008$ ), PSP-P ( $P=.002$ ) and PSP-PGF ( $P=.03$ ). \* $P<.05$ , \*\* $P<.01$ , \*\*\* $P<.001$ . Abbreviations: DMPI = dual-line midbrain PSP index; PSP = Progressive supranuclear palsy; PSP-CBS = PSP with predominant cortico-basal syndrome; PSP-F = PSP with predominant frontal presentation; PSP-SL = PSP with predominant speech/language disorder; PSP-P = PSP with predominant parkinsonism; PSP-PGF = PSP with progressive gait freezing.

**Figure S5.**



**Figure S5.** Logistic Regression model to distinguish participants with PSP from those with non-PSP parkinsonism in the whole cohort of 656 PSP and 802 non-PSP participants (Italian + international cohorts). The upper panel (A) shows the sigmoid with the probability of having PSP rather than non-PSP parkinsonian syndromes based on the linear predictor, and the corresponding mathematical function to calculate this probability in clinical practice using DMPI, age and sex of the subject. The bottom panel (B) shows the classification performance of the LR model in cross-validation in the whole cohort. In details, the model showed AUC: 0.95 (95%CI: 0.93, 0.97), and the other metrics were accuracy: 1315/1458, 90.19 (95%CI: 88.38, 92.00), sensitivity: 586/656, 89.33 (95%CI: 84.43, 94.22), specificity: 729/802, 90.90 (95%CI: 89.10, 98.06), calculated by using a discriminating probability threshold of 50% (probability  $\geq$  50% suggestive of PSP and probability  $<$  50% suggestive of non-PSP). Abbreviations: PSP = Progressive supranuclear palsy, DMPI = Dual-line midbrain PSP index; AUC = area under the curve.

**Figure S6.**



**Figure S6.** Logistic Regression model to distinguish participants with PSP from control participants in the whole cohort of 656 PSP and 610 control participants (Italian + international cohorts). The upper panel (A) shows the sigmoid with the probability of having PSP rather than being a control participant based on the linear predictor, and the corresponding mathematical function to calculate this probability in clinical practice using DMPI, age and sex of the subject. The bottom panel (B) shows the classification performance of the LR model in cross-validation in the whole cohort. In details, the model showed AUC: 0.96 (95%CI: 0.95, 0.97), and the other metrics were accuracy: 1158/1266, 91.47 (95%CI: 89.22, 93.71), sensitivity: 586/656, 89.33 (95%CI: 88.66, 90.00), specificity: 572/610, 93.77 (95%CI: 89.48, 98.06), calculated by using a discriminating probability threshold of 50% (probability  $\geq$  50% suggestive of PSP and probability  $<$  50% suggestive of control participant). Abbreviations: PSP = Progressive supranuclear palsy, DMPI = Dual-line midbrain PSP index; AUC = area under the curve.

## Appendix S1

In the current study, we analysed MRI data from multiple subject cohorts. MRI scans were acquired under various broader research protocols at the different institutions, within the context of the baseline visit procedures in clinical trials, or within specific data acquisition protocols for international data collection initiatives (i.e. ADNI, PPMI, 4RTNI), as described below. The MRI measurements investigated in this work were prospectively performed on the MR DICOM images between September 2024 and March 2025 for the specific aim of the current study.

### Participant cohorts

#### *University of Catanzaro (Italy)*

Participants (n=374) were consecutively recruited between January 2012 and June 2024 at the Movement Disorder Center of the University of Catanzaro, Italy. The cohort enrolled at this site included 136 participants with progressive supranuclear palsy (PSP), including 93 PSP-Richardson's syndrome (PSP-RS) and 43 PSP variants (35 PSP-parkinsonism [PSP-P], 3 PSP with predominant gait freezing [PSP-PGF], 3 PSP-corticobasal syndrome [PSP-CBS], 2 PSP-speech language [PSP-SL]), 30 MSA (17 MSA-cerebellar type [MSA-C], 13 MSA-parkinsonian type [MSA-P]), 208 Parkinson's disease (PD) and 85 control participants. The clinical diagnoses were performed by movement disorder specialists using international clinical diagnostic criteria (1-4). Participants with PSP enrolled before 2017 were diagnosed according to the National Institute of Neurological Disorders and Stroke and Society for PSP (NINDS- SPSP) criteria (5) and were retrospectively reclassified according to recent Movement Disorder Society (MDS) diagnostic criteria for possible or probable PSP (1). Control participants (n=85) were individuals aged above 50 years, independent in daily life activities, without clinical evidence of neurodegenerative or psychiatric disorders. All participants underwent a neurological examination including the MDS – sponsored revision of the Unified Parkinson's Disease Rating Scale part III (MDS-UPDRS-III)(6) in off-state and the Hoehn and Yahr (H-Y) rating scale, and the PSP rating scale (7) in participants with PSP. All study participants underwent a brain 3T MRI examination with either a 3T MR750 General Electric scanner (Discovery MR-750, GE, Milwaukee, WI, USA) with an 8-channel head coil (n = 407) or a hybrid 3T PET-MR scanner (Biograph mMR, Siemens Healthcare, Erlangen, Germany)

using a 16-channel PET-transparent head/neck coil ( $n = 52$ ; 8 PSP, 30 PD, 5 MSA, 9 control participants), including 3D T1-weighted images. For the MR-750 GE scanner the protocol included a 3D T1-weighted volumetric spoiled gradient echo (BRAVO, sagittal planes; repetition time/echo time 9.2/3.7 milliseconds; voxel size: 1.0 mm; slice thickness: 1.0 mm, frequency and phase encoding matrix  $256 \times 256$ ; flip angle  $12^\circ$ ; field of view 256 mm). For the SIEMENS scanner, the protocol included a 3D T1-weighted magnetization-prepared rapid acquisition gradient-echo sequence (MPRAGE, sagittal planes, phase encoding matrix  $256 \times 248$ , voxel size 1.0 mm, slice thickness: 1.0 mm. repetition time/echo time 2300/2.34 milliseconds, flip angle  $8^\circ$ , field of view 256 mm).

#### *Gosuranemab Trial (NCT03068468)*

In the NCT03068468 trial, participants with PSP were recruited at 90 centres across 13 countries (Australia, Austria, Canada, Germany, Spain, France, the United Kingdom, Greece, Italy, Japan, South Korea, Russia and the United States), between April 2017 and September 2019 (8). Participants were assigned to receive either gosuranemab (formerly BMS-986168 / IPN007 / BIIB092) or placebo for 52 weeks. Relevant inclusion criteria were: (i) age between 41-86 years and body weight 43–120 kg; (ii) history of postural instability or falls during the first 3 years from disease onset, vertical supranuclear gaze palsy or slow velocity of vertical saccades, and an akinetic-rigid syndrome; (iii) have PSP symptoms for less than 5 years; (iv) be able to ambulate independently or with limited assistance; (v) have a mini-mental state examination (MMSE) score of at least 20; (vi) live outside a nursing home or dementia care facility; (vii) no other notable neurological or psychiatric disorders including Alzheimer's disease, dementia with Lewy bodies, prion disease, Parkinson's disease, hydrocephalus or clinically relevant cerebrovascular disease. Available clinical information extracted from the study database included age, sex, age at onset, disease duration and PSP rating scale (7). MRI data were collected on 1.5T or 3T scanners with magnetization prepared 3D T1-weighted sequences  $1.0 \times 1.0 \times 1.2$  mm isotropic voxels acquired following the Alzheimer's Disease Neuroimaging Initiative (<https://adni.loni.usc.edu>) recommendations for volumetric analysis. The placebo arm of the trial included 158 participants with PSP with available MRI. One participant was excluded due to poor quality MR images, and

157 participants were included in the study. Among these, 153 had clinical scores consistent with the PSP-RS, while 4 participants had no falls or postural instability and were reclassified as PSP-P.

#### *Davunetide (AL-108-231) Trial (NCT01110720)*

In the NCT01110720 trial, participants with PSP were recruited at 48 centres across 6 countries (Australia, Canada, France, Germany, the United Kingdom, and the United States), between September 2010 and November 2012 (9). Participants were randomly assigned in a 1:1 ratio to davunetide or placebo for 52 weeks. Relevant inclusion criteria were: (i) age at disease onset between 41-85 years; (ii) at least a 12-month history of postural instability or falls during the first 3 years from disease onset, supranuclear ophthalmoplegia or reduced downward saccade velocity, and prominent axial rigidity; (iii) have PSP symptoms for either less than 5 years, or more than 5 years with a PSPRS score or no more than 40; (iv) be able to ambulate independently (or walk 5 steps with minimal assistance); (v) have a MMSE score of at least 15; (vi) live outside a nursing home or dementia care facility. Available clinical information extracted from the study database included age, sex and PSP rating scale (7). MRI data were collected on 1.5 or 3T scanners with magnetization prepared 3D T1-weighted sequences 1.0 x 1.0 x 1.0 mm isotropic voxels acquired following the Alzheimer's Disease Neuroimaging Initiative (<https://adni.loni.usc.edu>) recommendations for volumetric analysis. The placebo arm of the trial included 144 participants with PSP-RS, and all of them were included in this study.

#### *Tideglusib Trial (NCT01049399)*

In the NCT01049399 trial, participants with PSP were recruited at 24 centres across 4 countries (Germany, Spain, the United Kingdom and the United States), between December 2009 and November 2011 (10). Participants were randomly assigned to tideglusib 600 mg, tideglusib 800 mg or placebo with a 2:2:1 ratio for 52 weeks. Relevant inclusion criteria were: (i) age at disease onset between 40 and 85 years; (ii) fulfilling possible or probable NINDS- SPSP criteria; (iii) Mild-to-moderate stage of disease severity according to score of 1 to 4 in Golbe Staging System; (iv) Brain MRI examination within 24 months before baseline visit excluding other potential causes of parkinsonism, especially cerebrovascular lesions and space occupying

lesions. Available clinical information extracted from the study database included age, sex, age at onset, disease duration and PSP rating scale (7). MRI data were collected on 1.5 or 3T scanners with magnetization prepared 3D T1-weighted sequences 1.0 x 1.0 x 1.0 mm isotropic voxels acquired following the Alzheimer's Disease Neuroimaging Initiative (<https://adni.loni.usc.edu>) recommendations for volumetric analysis. The placebo arm of the trial included 32 participants with PSP, of whom 9 had available brain MRI (4 PSP-RS and 5 PSP-P). All of them were included in this study.

#### *DescribePSP Network cohort*

DescribePSP (DZNE Clinical Register Study of Neurodegenerative Disorders - PSP) is a large German multicentre research network set up in 2015 organized by the German Centre for Neurodegenerative Diseases (DZNE), prospectively collecting comprehensive clinical data, imaging data and biomaterials of participants with PSP (11). Participants with a clinical diagnosis of PSP according to the MDS diagnostic criteria (1) were consecutively enrolled in the observational DescribePSP study at 11 tertiary care centres with expertise in movement disorders, in Berlin, Bonn, Dresden, Gottingen, Greifswald, Hanover, Cologne, Magdeburg, Munich, Rostock, and Tübingen. Available clinical information included age, sex, age at onset, disease duration and PSP rating scale (7). MRI data were collected on 3T scanners acquired following the Alzheimer's Disease Neuroimaging Initiative (<https://adni.loni.usc.edu>) recommendations for volumetric analysis. As per data collection in August 2022, the cohort included 37 participants with PSP with available MRI, who were enrolled between August 2016 and October 2021. Among these, 5 participants only fulfilled criteria for “suggestive of PSP” condition and were excluded (in the “diagnostic concerns” category in Figure 1. All the remaining 32 participants were included in this study.

#### *PROSPECT-UK cohort*

The PROSPECT study natural history cohort is an observational longitudinal study enrolling participants at 7 UK sites (University College London [UCL], Oxford, Cambridge, Newcastle, Brighton, Newport, and Manchester), starting in 2015 (12). Participants entering the study were diagnosed as PSP following the

NINDS-SPSP criteria (5) and reclassified at the end of baseline recruitment according to current MDS PSP criteria (1). All reclassified PSP cases fulfilled at least “possible” diagnostic criteria and were stratified into PSP-RS, PSP-subcortical, and PSP-cortical groups. The PSP-subcortical group included cases with PSP-P, PSP- with progressive gait freezing (PSP-PGF), and PSP-oculomotor (PSP-OM); the PSP-cortical group included PSP with predominant cortico-basal presentation (PSP-CBS) and PSP with predominant frontal presentation (PSP-F). Participants with CBS were diagnosed following the Armstrong criteria (4). Participants with MSA were diagnosed following the revised Gilman criteria and stratified into MSA-C and MSA-P subgroups (3). Participants with progressive movement or cognitive disorders, thought to have atypical parkinsonian syndromes but not meeting any of the above diagnostic criteria, were classified as indeterminate (IDT) cases. Available clinical data included age, sex, age at onset, disease duration and appropriate clinical scores (PSP rating scale, MDS-UPDRS-III scale, Unified Multiple System Atrophy Rating Scale [UMSARS] (13)) were extracted from the study dataset. PROSPECT participants underwent brain MRI with volumetric T1-weighted images on 3T scanners (Siemens, Prisma, or TRIO) at UCL, Cambridge or Oxford center. Scan protocols were designed based on the international Genetic Frontotemporal Dementia Initiative protocols (MPRAGE, repetition time: 2 s, echo time: 2.93 ms, Flip angle 8°, 1.1mm isotropic). A subset of participants with CBS underwent cerebrospinal fluid (CSF) analysis, and AD biomarkers including CSF total tau (T-tau) and  $\beta$ -amyloid 1-42 ( $A\beta$ 1-42) levels (INNOTEST ELISA – Fujirebio Europe N.V., Gent, Belgium) were investigated at the UK Dementia Research Institute Fluid Biomarker Laboratory, University College London, London, UK. Participants with CBS showing a T-tau /  $A\beta$ 1-42 ratio  $> 1$  were defined as having CBS with likely underlying AD pathologic features (CBS-AD), and those with this ratio  $< 1$  were defined as CBD-4RT because they are likely to have underlying 4R tauopathy (12). Data of all participants with available brain MRI were queried to PROSPECT investigators and obtained in February 2024. The whole dataset included 89 participants with available brain MRI, who were enrolled between July 2015 and May 2019. Among these, four participants were excluded because of missing basic demographic and clinical information, 9 participants were excluded because of a IDT diagnosis, 3 participants were excluded for diagnostic concerns based on conflicting information and 3 participants were excluded because of poor quality MRI with motion artifacts. Thus, 70 participants were included in this study, stratified as follows: 23 PSP-RS, 14 PSP variants (8 cortical

and 6 subcortical), 17 MSA (6 MSA-C, 8 MSA-P and 3 MSA-indeterminate) and 16 CBS participants (4 CBS-AD, 4 CBS-4RT and 8 CBS-indeterminate).

#### *University of Athens cohort*

Participants (n=196) were consecutively recruited at the University of Athens, Greece between January 2014 and December 2023. The clinical diagnoses were performed by movement disorder specialists using international clinical diagnostic criteria (1-4). The cohort enrolled at this site included 46 participants with PSP-RS, 24 PSP variants (11 PSP-CBS, 7 PSP-F, 4 PSP-SL, 1 PSP-P, 1 PSP-PGF), 40 MSA (24 MSA-C, 16 MSA-P), 43 CBS and 81 control participants. All PSP cases fulfilled possible or probable diagnostic criteria for PSP (1). Control participants (n=81) were individuals aged  $\geq 50$  years, with no history of neurological, psychiatric, or other major disease and no signs of parkinsonism or cognitive dysfunction. All participants underwent a neurological examination including the Unified Parkinson's Disease Rating Scale part III (UPDRS-III) (14). The UPDRS-III scores extracted from the study dataset were converted into the MDS-UPDRS-III scores as previously suggested (15) to merge data with those from other cohorts. Moreover, the PSP rating scale (7) was employed to evaluate participants with PSP and the UMSARS (13) to evaluate those with MSA diagnosis. All participants underwent a brain MRI examination either at our institution with a Philips Medical Systems Achieva 3.0 T scanner (233 subjects) or at other sites with 1.5T or 3T scanners (26 subjects with 1.5T and 18 subjects with 3T). A subset of participants with CBS underwent CSF analysis, and AD biomarkers including CSF total tau (T-tau), Phospho-Tau 181 (pTau181), beta-Amyloid 1-40 (A $\beta$ 40) and 1-42 (A $\beta$ 42) levels were measured by double sandwich, enzyme-linked immunosorbent assay (ELISA) in duplicate with commercially available kits ELISA kits from EUROIMMUN, based on the manufacturer's instructions, using the EUROIMMUN Analyzer I (EUROIMMUN, Medizinische Labordiagnostika AG, Lübeck, Germany), as previously described (16). CBS participants with abnormal A $\beta$ 42, A $\beta$ 42/A $\beta$ 40, and phospho-Tau levels were considered to have concomitant underlying AD pathology (CBS-AD). The following local cut-off values of the Neurochemistry and Biomarkers Unit were employed (A $\beta$ 42 < 480 pg/mL; total Tau > 400 pg/mL, phospho-Tau > 60 pg/mL A $\beta$ 42/A $\beta$ 40 ratio < 0.094), as previously reported (16-17).

#### *University of Turku cohort*

This cohort included a retrospective set of 20 individuals with pathologically proven PSP (n=8), MSA (n=5) or PD (n=7), who underwent clinical assessment and brain MRI between June 2006 and November 2019. Available data extracted from medical charts included age, sex, age at onset and disease duration at MRI. Three participants were excluded due to severe motion artifacts on MR images, thus the final cohort included 7 PSP, 5 MSA and 5 PD participants.

#### *4R-Tau Neuroimaging Initiative (4RTNI) cohort*

The 4R-Tau Neuroimaging Initiative, 4RTNI (details at <https://4rtni-ftldni.ini.usc.edu/> and <https://ida.loni.usc.edu/home/projectPage.jsp?project=4RTNI>) is an observational American study enrolling participants with Progressive Supranuclear Palsy or Cortico-basal Syndrome (CBS) aged between 45 and 90 years old. It was funded through the National Institute of Aging and The Tau Research Consortium. The primary goal of 4RTNI was to identify neuroimaging and biomarker indicators for disease progression in the 4-repeat tauopathy neurodegenerative diseases, progressive supranuclear palsy (PSP) and cortico-basal degeneration (CBD). Participants with PSP were diagnosed according to the possible or probable National Institute of Neurological Disorders and Stroke and Society for PSP (NINDS- SPSP) criteria (5) and were classified as PSP-RS. Available clinical data extracted from the study dataset included age, sex, age at onset, disease duration and PSP rating scale (7). Participants underwent brain MRI with 3T scanners (GE Signa HDxt, GE Discovery MR750 or SIEMENS TrioTim) with the protocol including T1-weighted 3D images (sagittal planes, slice thickness 1.0 for SIEMENS and 1.2 for GE scanners). A subgroup of participants underwent post-mortem pathological diagnostic confirmation. The data is the result of collaborative efforts at two sites in North America: San Francisco, California, USA and Baltimore, Maryland, USA. (<https://clinicaltrials.gov/study/NCT01804452>).

All available data of 4RTNI participants at baseline were selected, resulting in 127 subjects. After excluding subjects with “other” diagnosis, 120 participants (72 PSP and 48 CBS) were identified. Among these subjects, 4 participants with PSP had no available MRI and 2 participants with PSP had poor quality MRI, not allowing to perform reliable measurements. Among the participants with CBS, 9 subjects had high ( $\geq 5/16$ ) PSP rating

scale ocular scores, reflecting at least mild ocular motor dysfunction, thus were re-classified as PSP-CBS, in agreement with the most recent diagnostic criteria for PSP, which operationalized the diagnosis of several PSP variants (1). Thus, a total of 75 participants with PSP (66 PSP-RS and 9 PSP variants) and 39 participants with CBS enrolled between March 2011 and September 2014 were included in this study. No information on Alzheimer's disease biomarkers in CBS participants were available for this cohort.

#### *Parkinson's Progression Markers Initiative (PPMI) cohort*

The PPMI is an ongoing, observational, international study aimed at identifying blood, cerebrospinal fluid, genetic, and imaging biomarkers for PD progression (18). PD and control participants at baseline visit aged  $\geq 50$  years with available 3T brain MRI and 3D T1-weighted images acquired on sagittal plane with voxel size and slice thickness of 1.0 mm were selected from the PPMI dataset. Available clinical information extracted from the PPMI dataset included age, sex, age at disease onset, disease duration, MDS-UPDRS-III score performed in OFF- state and H-Y score. Detailed information on the PPMI MRI protocol is available at <http://www.ppmi-info.org>. Data were downloaded from the PPMI database in July 2024, resulting in 375 participants with PD and 127 control subjects fulfilling inclusion criteria, enrolled between July 2012 and May 2024. Nine participants with PD were excluded due to diagnostic concerns (diagnosis varied across visits), thus the PPMI cohort included 366 PD and 127 control participants. In addition, all participants with available neuropathological post-mortem diagnostic confirmation and brain MRI were selected, resulting in 24 subjects. Two participants were excluded due to alternative diagnoses (one had cerebrovascular disease with no Lewy body pathology, and another had no brain diseases). Thus, 22 participants were included (21 with a pathologically proven PD and one with a neuropathological diagnosis of MSA).

#### *Alzheimer's Disease Neuroimaging Initiative (ADNI) cohort*

The ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD (19). The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological

assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD). Control participants at baseline visit aged  $\geq 50$  years with available 3T brain MRI and 3D T1-weighted images acquired on sagittal plane with voxel size and slice thickness of 1.0 mm were selected from the ADNI dataset. Detailed information on the ADNI MRI protocol is available at <https://adni.loni.usc.edu>. Data were downloaded from the ADNI database in July 2024, resulting in 202 control participants fulfilling inclusion criteria, who were enrolled between August 2006 and September 2024.

#### *Open Access Series of Imaging Studies - 3 (OASIS-3) cohort*

OASIS-3 is a retrospective compilation of data for 1378 participants that were collected across several ongoing projects through the WUSTL Knight ADRC over the course of 30 years, including brain MRI and PET data (20). Participants include 755 cognitively normal adults and 622 individuals at various stages of cognitive decline ranging in age from 42-95yrs. A total of 115 control participants aged  $\geq 50$  years with available 3T brain MRI and 3D T1-weighted images were randomly selected from the OASIS-3 dataset and included in the study. These subjects were enrolled in OASIS-3 between October 2017 and February 2022.

#### **Statistical analysis**

Data normality was assessed using the Shapiro-Wilk test. Demographic and clinical information (age, sex, disease severity) were compared using Fisher's test, analysis of variance (ANOVA), Kruskal-Wallis test, t-test or Wilcoxon rank sum test, as appropriate. Imaging data (A, B and DMPI values) were compared across groups using ANCOVA with age and sex as covariates. In all primary (PSP vs non-PSP and PSP vs control participants) and secondary comparisons (pairwise comparison among all study subgroups), p values were corrected for the number of tests in the analysis (Bonferroni correction). Associations between imaging and clinical data were investigated using multivariate linear regression models including age and sex. All tests were two-tailed, with the significance threshold set at  $\alpha = 0.05$ . P values were corrected according to Bonferroni. The percentage difference for A and B midbrain measures in participants with PSP was calculated as follows:

$$\%diffx = (x_{PSP} - \mu_{xControls}) / \mu_{xControls}$$

Where  $x$  was the measure of interest,  $x_{PSP}$  was the  $x$  value of each individual PSP participant and  $\mu_{Controls}$  was the mean of  $x$  in the control participant group. The coefficient of variation for imaging measures was calculated as the ratio between its standard deviation and its mean in the PSP group, as follows:

$$Coeff.varx = \sigma_{xPSP} / \mu_{xPSP}$$

Where  $x$  was the measure of interest,  $\sigma_{xPSP}$  was the standard deviation of  $x$  in the PSP group and  $\mu_{xPSP}$  was the mean of  $x$  in the PSP group.

To evaluate the reliability of the proposed manual midbrain measurement, intra-rater and inter-rater reliability was assessed. Intraclass correlation coefficients (ICCs) were calculated both for inter-rater and intra-rater reliability (one rater performed the measurements twice with a time interval of around two weeks) using 2-way random-effects ICCs for absolute agreement based on single measurements, in a subset of 150 MR images (50 PSP, 50 PD, 50 control participants).

The performance of DMPI values (without considering age and sex) in distinguishing participants with PSP from those with non-PSP parkinsonism and control subjects were investigated using standard Receiver Operating Characteristic (ROC) analysis, as commonly performed in biomarker studies. Optimal cut-offs, defined as the values with the highest sum of sensitivity and specificity (Youden's method), and 95% confidence intervals, were calculated using R pROC software package with bootstrapping (n=2,000 iterations). This method was also used to investigate the performance of DMPI and other MRI planimetric or linear measures in distinguishing participants with PSP from those with non-PSP parkinsonism in a subset of the study cohort (see paragraph below). Reliability and performance metrics were evaluated as follows:  $\geq 0.90$  Excellent, 0.75–0.89 Good, 0.60–0.74 Moderate,  $< 0.60$  Poor). Statistical analyses were performed using R statistical software (R for Unix/Linux, version 4.0.4, the R Foundation for Statistical Computing, 2021).

*Comparison between the DMPI and other planimetric or linear measures in distinguishing participants with PSP from those with non-PSP parkinsonism.*

We compared the DMPI with other previously described measures, including a midbrain line, midbrain area, midbrain-to-pons (M/P) area ratio, Magnetic Resonance Parkinsonism Index (MRPI) and MRPI 2.0. All these

measures were performed in a sub-cohort of 161 PSP and 203 non-PSP participants, which was created by randomly selecting around 25% of individuals from each participant group within the main cohort, maintaining the original balance between Italian and international datasets. The midbrain line was performed manually by the same neuroradiology technician who performed the DMPI measures in all study participants (I.C.). It was measured on midsagittal T1-weighted MR images as the maximum anteroposterior midbrain diameter, including the quadrigeminal plate, according to the original description (21). The planimetric measures (midbrain area, M/P area ratio, MRPI, MRPI 2.0) were performed automatically using a previously described in-house algorithm (22-25). The algorithm was previously developed using R2017a MATLAB software, and it is based on several consecutive operator-independent procedural steps (22-25). In brief, T1-weighted structural MRI images are normalized into Montreal Neurological Institute (MNI) template using FSL software (FMRIB Software Library). Intensity normalization is performed with FreeSurfer software package. The algorithm identifies automatically the mid-sagittal plane using anatomical landmarks as the slice with the maximal expansion of the Sylvius aqueduct, and the midbrain and pons are automatically segmented on the mid-sagittal image to obtain midbrain area and pons area (22-23). Parasagittal images exposing the left and right middle cerebellar peduncles are generated and the MCP width is automatically measured on consecutive sagittal slices, and averaged. A volumetric slab of 40 mm (0.5-mm section thickness) tangent to the floor of the fourth ventricle is generated to expose the SCPs, and the left-to-right width of both SCP is automatically measured on two consecutive images, and averaged (22-24). Finally, reformatted volumetric slab (including 35 slices each with 1 mm thickness) parallel to the subcallosal line is automatically generated to expose several axial views of the third ventricle and the frontal horns of the lateral ventricles. The algorithm identifies the slice with the largest third ventricle expansion and calculates the third ventricle width as the mean of six automated linear measures. The frontal horns' width is measured as the largest left-to-right distance between the lateral borders of frontal horns on 15 consecutive axial slices, and the largest measure is selected (25).

MRPI and MRPI 2.0 indexes are calculated as follows:

$$MRPI = (pons\ area / midbrain\ area) * (MCP\ width / SCP\ width) \quad (24)$$

$$MRPI\ 2.0 = MRPI * (third\ ventricle\ width / frontal\ horns'\ width) \quad (25)$$

The performance and 95% CIs of each marker was investigated in differentiating participants with PSP from those with non-PSP parkinsonism using ROC analysis with bootstrapping (n=2,000 iterations), using the pROC package in R software. Confidence intervals were calculated using DeLong method. The classification performances between the DMPI and each other planimetric or linear marker were compared using the De Long test, in the participant group where both the DMPI and the other considered marker was available (paired test). Optimal cut-offs were identified as the values with the highest sum of sensitivity and specificity (Youden method). Beyond ROC performances, we employed a two-cutoffs approach and compared the percentage of participants with measures in a possible “grey zone” for each marker (defined as the range between the cut-off corresponding to 95% sensitivity and that corresponding to 95% specificity), as common practice in recent biomarker studies (26-27). A smaller percentage of cases in the grey zone means a lower number of uncertain results, and thus a more powerful marker.

### *Logistic regression models*

The main outcome measure was the classification performance of the DMPI in distinguishing participants with PSP from those with non-PSP parkinsonism and control participants. To this aim, we employed a multivariate logistic regression (LR) classifier including the DMPI value as predictor, and confounding factors (age and sex). Age and sex were included since these variables showed significant associations with imaging midbrain measures. Separate models for the PSP vs non-PSP and for the PSP vs control participant comparisons, the former to provide accuracy of the model in a clinical scenario (where control participants are typically absent), and the second as “reference” comparison in distinguishing the disease of interest from healthy subjects. The LR models were trained using standard default parameters to calculate the intercept and the predictors’ coefficients to maximize classification accuracy, and the models’ performances were then evaluated using area under the ROC curve (AUC), accuracy, sensitivity, specificity and positive and negative predictive value metrics in cross-validation or in independent test sets. Confidence intervals for AUC in cross-validation were calculated as follows:

$$95\% \text{ CI} = m - 1.96 \times s/\sqrt{N}, m + 1.96 \times s/\sqrt{N}$$

where  $N$  is the number of cross-validation iterations, and  $m$  and  $s$  are the mean and the standard deviation of AUC values, respectively. A common practice to calculate the performance of classifiers is to identify the optimal cut-off (i.e. by Youden index); the "optimal cut-off", often variable across cohorts, however, might be not easily interpretable in this context. The LR model assign to each participant an individual probability score of having PSP rather than the other condition (non-PSP or control participants); thus, we employed a more intuitive approach by using a fixed discriminating probability threshold (cut-off) of 50% in all study analyses to calculate sensitivity, specificity and accuracy metrics (where a score  $\geq 50\%$  means that the subject is more likely to have PSP, and a score below this threshold is suggestive of non-PSP or control participant). First, we performed a stratified 5-fold cross-validation (repeated five times) to investigate the performance in the training set (Italian cohort,  $n=459$ ; 136 PSP, 238 non-PSP 85 control participants). Hyperparameter optimization was performed in cross-validation procedure using *RandomizedSearchCV*, involving the following parameters: exploring regularization strength (C) [0.001, 0.01, 0.1, 1, 10, 100], penalty type [l1 or l2], and solver [liblinear or saga]. Class weight was set to "balanced," and the maximum number of iterations was fixed at 100,000 to guarantee convergence. Subsequently, we validated the performance of the best model in the independent external test set (international cohort,  $n=1609$ ; 520 PSP, 564 non-PSP, 525 control participants).

Then, to build a reliable classifier to be used in routine practice, we leveraged our whole participant cohort (Italian + international cohort: 656 PSP, 802 non-PSP, 610 control participants); the classification performance of the LR models for "PSP vs non-PSP" and "PSP vs control participants" comparisons in this large international cohort were assessed through cross-validation procedure, as described above. In addition, this model was tested in a small independent external test set of participants ( $n=43$ ) with pathologically proven diagnoses of PSP, PD or MSA. The LR model based on DMPI, age and sex assigned to each participant the probability score of having PSP rather than a non-PSP parkinsonian syndrome (or control participants). The individual probability score could be easily calculated using the regression-derived mathematical functions for the PSP vs non-PSP or the PSP vs control participant LR classifiers, which are provided in the results. Moreover, to further simplify the use of this tool in routine imaging examination, we developed a calculator based on the PSP vs non-PSP model available free of charge on github, at [https://neuroimagingunicz.github.io/mri\\_calc/](https://neuroimagingunicz.github.io/mri_calc/) website.

The classification performance of LR models based on DMPI, age and sex in distinguishing PSP from non-PSP parkinsonism were then investigated in a sub-cohort of participants at the early stage of the diseases (within 1, 2 or 3 years from the disease onset), in cross-validation. Finally, comprehensive analyses were performed to calculate DMPI classification performances (in LR models) in distinguishing pairwise across all participant subgroups. To address class imbalance due to some groups having a sample size significantly larger than others, we employed a cluster-based under-sampling strategy to create balanced datasets while preserving the diversity and distribution of the original data. This approach involved clustering the larger class into representative subsets and subsequently sampling a fixed number of subjects from each cluster. This approach ensured that the model performance was not biased toward the overrepresented class, resulting in more reliable results. The under-sampling strategy was repeated ten times and the average classification performances in each comparison were calculated through cross-validation procedure. All logistic regression model analyses were conducted in Python 3.9 using the scikit-learn library (version 1.0.1).

## **Appendix S2: Supplementary Results**

### *Dual-line Midbrain PSP Index (DMPI)*

We traced two different linear midbrain measurement (Figure 1): the line A to assess midbrain body atrophy, and the line B to assess the atrophy of the anterior midbrain part extending into the mesencephalic beak. This new measure ( $[A+B]/2$ ) was termed Dual-line Midbrain PSP Index (DMPI) and showed a coefficient of variation in PSP only slightly higher than A (0.19 vs 0.16). No associations were found between DMPI values and MRI field strength (1.5T/3T). Conversely, midbrain measurements were significantly associated with age and sex, thus we employed logistic regression (LR) models including age and sex to investigate classification performances in distinguishing between participant groups.

### *Midbrain linear measurement errors and classification performances*

In the current study, we proposed a marker calculated averaging two lines (DMPI). Midbrain linear measurements are performed manually and typically just a few millimetres in length, thus small measurement errors in the magnitude of around 1 mm may affect individual subject classification accuracy; in this context, we hypothesized that a marker based on averaging two measures might be less sensitive a small measurement mistakes than one focusing on a single linear width. The decision of using the average of A and B lines rather than any other formula was not the result of several tries, rather was decided before the analyses. To test our hypothesis, we assessed the impact of possible measurement error on midbrain measures, by artificially modifying A and B measures by 10% or 20% in both PSP-RS and PD groups, to account for the presence of possible deleterious mistakes. Then we investigated how these mistakes affected the classification accuracy of each measure (A, B and their average [DMPI]) in differentiating participant with PSP from those with non-PSP parkinsonism in the whole cohort of 656 PSP and 802 non-PSP participants. In detail, we hypothesized three scenarios where measurement mistakes could have a deleterious effect on classification accuracy: (i) larger measures in participants with PSP, (ii) smaller measures in participants with non-PSP parkinsonism, and (iii) a combination of both these errors. Thus, actual A or B values were increased in PSP and/or decreased in non-PSP by 10% or 20% magnitude. These changes were performed alternatively either for A or B measure, and the DMPI was calculated after measurement adjustment. In all cases, the classification performances were assessed by logistic regression analysis including the midbrain measure, age and sex. The DMPI showed the best performance in all scenarios, demonstrating the advantage of using a dual-line measurement, whose classification accuracy was less affected by possible measurement mistakes (Table S2).

## References

1. Höglinger GU, Respondek G, Stamelou M, Kurz C, Josephs KA, Lang AE, Mollenhauer B, Müller U, Nilsson C, Whitwell JL, Arzberger T, Englund E, Gelpi E, Giese A, Irwin DJ, Meissner WG, Pantelyat A, Rajput A, van Swieten JC, Troakes C, Antonini A, Bhatia KP, Bordelon Y, Compta Y, Corvol JC, Colosimo C, Dickson DW, Dodel R, Ferguson L, Grossman M, Kassubek J, Krismer F, Levin J, Lorenzl S, Morris HR, Nestor P, Oertel WH, Poewe W, Rabinovici G, Rowe JB, Schellenberg GD, Seppi K, van Eimeren T, Wenning GK, Boxer AL, Golbe LI, Litvan I; Movement Disorder Society-endorsed PSP Study Group. Clinical diagnosis of progressive supranuclear palsy: The movement disorder society criteria. *Mov Disord*. 2017 Jun;32(6):853-864. doi: 10.1002/mds.26987.
2. Postuma RB, Berg D, Stern M, Poewe W, Olanow CW, Oertel W, Obeso J, Marek K, Litvan I, Lang AE, Halliday G, Goetz CG, Gasser T, Dubois B, Chan P, Bloem BR, Adler CH, Deuschl G. MDS clinical diagnostic criteria for Parkinson's disease. *Mov Disord*. 2015 Oct;30(12):1591-601. doi: 10.1002/mds.26424.
3. Gilman S, Wenning GK, Low PA, Brooks DJ, Mathias CJ, Trojanowski JQ, Wood NW, Colosimo C, Dürr A, Fowler CJ, Kaufmann H, Klockgether T, Lees A, Poewe W, Quinn N, Revesz T, Robertson D, Sandroni P, Seppi K, Vidailhet M. Second consensus statement on the diagnosis of multiple system atrophy. *Neurology*. 2008 Aug 26;71(9):670-6. doi: 10.1212/01.wnl.0000324625.00404.15.
4. Armstrong MJ, Litvan I, Lang AE, et al. Criteria for the diagnosis of corticobasal degeneration. *Neurology*. 2013;80(5):496-503. doi:10.1212/WNL.0b013e31827f0fd1.
5. Litvan I, Agid Y, Calne D, et al. Clinical research criteria for the diagnosis of progressive supranuclear palsy (Steele-Richardson-Olszewski syndrome): report of the NINDS-SPSP international workshop. *Neurology*. 1996;47(1):1-9. doi:10.1212/WNL.47.1.1.
6. Goetz CG, Tilley BC, Shaftman SR, Stebbins GT, Fahn S, Martinez-Martin P, Poewe W, Sampaio C, Stern MB, Dodel R, Dubois B, Holloway R, Jankovic J, Kulisevsky J, Lang AE, Lees A, Leurgans S, LeWitt PA, Nyenhuis D, Olanow CW, Rascol O, Schrag A, Teresi JA, van Hilten JJ, LaPelle N; Movement Disorder Society UPDRS Revision Task Force. Movement Disorder Society-sponsored revision of the Unified Parkinson's Disease Rating Scale (MDS-UPDRS): scale presentation and clinimetric testing results. *Mov Disord*. 2008 Nov 15;23(15):2129-70. doi: 10.1002/mds.22340.

7. Golbe LI, Ohman-Strickland PA. A clinical rating scale for progressive supranuclear palsy. *Brain*. 2007 Jun;130(Pt 6):1552-65. doi: 10.1093/brain/awm032.
8. Dam T, Boxer AL, Golbe LI, et al; PASSPORT Study Group. Safety and efficacy of anti-tau monoclonal antibody gosuranemab in progressive supranuclear palsy: a phase 2, randomized, placebo-controlled trial. *Nat Med*. 2021; 27(8):1451-1457. doi: 10.1038/s41591-021-01455-x.
9. Boxer AL, Lang AE, Grossman M, et al; AL-108-231 Investigators. Davunetide in patients with progressive supranuclear palsy: a randomised, double-blind, placebo-controlled phase 2/3 trial. *Lancet Neurol*. 2014; 13(7):676-85. doi: 10.1016/S1474-4422(14)70088-2.
10. Höglinger GU, Huppertz HJ, Wagenpfeil S, et al; TAUROS MRI Investigators. Tideglusib reduces progression of brain atrophy in progressive supranuclear palsy in a randomized trial. *Mov Disord*. 2014; 29(4):479-87. doi: 10.1002/mds.25815.
11. Respondek G, Höglinger GU. DescribePSP and ProPSP: German Multicenter Networks for Standardized Prospective Collection of Clinical Data, Imaging Data, and Biomaterials of Patients With Progressive Supranuclear Palsy. *Front Neurol*. 2021; 12:644064. doi: 10.3389/fneur.2021.644064.
12. Jabbari E, Holland N, Chelban V, et al. Diagnosis Across the Spectrum of Progressive Supranuclear Palsy and Corticobasal Syndrome. *JAMA Neurol*. 2020 Mar 1;77(3):377-387. doi: 10.1001/jamaneurol.2019.4347.
13. Wenning GK, Tison F, Seppi K, Sampaio C, Diem A, Yekhelef F, Ghorayeb I, Ory F, Galitzky M, Scaravilli T, Bozi M, Colosimo C, Gilman S, Shults CW, Quinn NP, Rascol O, Poewe W; Multiple System Atrophy Study Group. Development and validation of the Unified Multiple System Atrophy Rating Scale (UMSARS). *Mov Disord*. 2004 Dec;19(12):1391-402. doi: 10.1002/mds.20255.
14. Fahn S, Elton RL, UPDRS program members. Unified Parkinson's Disease Rating Scale. In: S Fahn, CD Marsden, M Goldstein, DB Calne, editors. *Recent Developments in Parkinson's Disease*, vol. 2. Florham Park, NJ: Macmillan Healthcare Information; 1987. p 153–163, 293–304.
15. Hentz JG, Mehta SH, Shill HA, Driver-Dunckley E, Beach TG, Adler CH. Simplified conversion method for unified Parkinson's disease rating scale motor examinations. *Mov Disord*. 2015 Dec;30(14):1967-70. doi: 10.1002/mds.26435.

16. Constantinides VC, Tentolouris-Piperas V, Paraskevas GP, Pyrgelis ES, Velonakis G, Karavasilis E, Toulas P, Boufidou F, Stefanis L, Kapaki E. Hippocampal subfield volumetry in corticobasal syndrome of diverse underlying pathologies. *J Neurol*. 2023 Apr;270(4):2059-2068. doi: 10.1007/s00415-022-11538-5.
17. Paraskevas GP, Kasselimis D, Kourtidou E, Constantinides V, Bougea A, Potagas C, Evdokimidis I, Kapaki E. Cerebrospinal Fluid Biomarkers as a Diagnostic Tool of the Underlying Pathology of Primary Progressive Aphasia. *J Alzheimers Dis*. 2017;55(4):1453-1461. doi: 10.3233/JAD-160494.
18. Marek K, Chowdhury S, Siderowf A, Lasch S, Coffey CS, Caspell-Garcia C, Simuni T, Jennings D, Tanner CM, Trojanowski JQ, Shaw LM, Seibyl J, Schuff N, Singleton A, Kieburtz K, Toga AW, Mollenhauer B, Galasko D, Chahine LM, Weintraub D, Foroud T, Tosun-Turgut D, Poston K, Arnedo V, Frasier M, Sherer T; Parkinson's Progression Markers Initiative. The Parkinson's progression markers initiative (PPMI) - establishing a PD biomarker cohort. *Ann Clin Transl Neurol*. 2018 Oct 31;5(12):1460-1477. doi: 10.1002/acn3.644.
19. Petersen RC, Aisen PS, Beckett LA, Donohue MC, Gamst AC, Harvey DJ, Jack CR Jr, Jagust WJ, Shaw LM, Toga AW, Trojanowski JQ, Weiner MW. Alzheimer's Disease Neuroimaging Initiative (ADNI): clinical characterization. *Neurology*. 2010 Jan 19;74(3):201-9. doi: 10.1212/WNL.0b013e3181cb3e25.
20. OASIS-3: Longitudinal Neuroimaging, Clinical, and Cognitive Dataset for Normal Aging and Alzheimer Disease Pamela J LaMontagne, Tammie L.S. Benzinger, John C. Morris, Sarah Keefe, Russ Hornbeck, Chengjie Xiong, Elizabeth Grant, Jason Hassenstab, Krista Moulder, Andrei Vlassenko, Marcus E. Raichle, Carlos Cruchaga, Daniel Marcus, 2019. medRxiv. Doi: 10.1101/2019.12.13.19014902.
21. Warmuth-Metz M, Naumann M, Csoti I, Solymosi L. Measurement of the midbrain diameter on routine magnetic resonance imaging: a simple and accurate method of differentiating between Parkinson disease and progressive supranuclear palsy. *Arch Neurol*. 2001 Jul;58(7):1076-9. doi: 10.1001/archneur.58.7.1076.
22. Nigro S, Arabia G, Antonini A, et al. Magnetic Resonance Parkinsonism Index: diagnostic accuracy of a fully automated algorithm in comparison with the manual measurement in a large Italian

- multicentre study in patients with progressive supranuclear palsy. *Eur Radiol.* 2017; 27(6):2665-2675. doi: 10.1007/s00330-016-4622-x.
23. Nigro S, Cerasa A, Zito G, et al. Fully automated segmentation of the pons and midbrain using human T1 MR brain images. *PLoS One.* 2014; 9(1):e85618. doi: 10.1371/journal.pone.0085618.
24. Nigro S, Antonini A, Vaillancourt DE, et al. Automated MRI Classification in Progressive Supranuclear Palsy: A Large International Cohort Study. *Mov Disord.* 2020; 35(6):976-983. doi: 10.1002/mds.28007.
25. Quattrone A, Bianco MG, Antonini A, et al. Development and Validation of Automated Magnetic Resonance Parkinsonism Index 2.0 to Distinguish Progressive Supranuclear Palsy-Parkinsonism From Parkinson's Disease. *Mov Disord.* 2022; 37(6):1272-1281. doi: 10.1002/mds.28992.
26. Barthélemy NR, Salvadó G, Schindler SE, He Y, Janelidze S, Collij LE, Saef B, Henson RL, Chen CD, Gordon BA, Li Y, La Joie R, Benzinger TLS, Morris JC, Mattsson-Carlsson N, Palmqvist S, Ossenkoppele R, Rabinovici GD, Stomrud E, Bateman RJ, Hansson O. Highly accurate blood test for Alzheimer's disease is similar or superior to clinical cerebrospinal fluid tests. *Nat Med.* 2024 Apr;30(4):1085-1095. doi: 10.1038/s41591-024-02869-z.
27. Wang J, Huang S, Lan G, Lai YJ, Wang QH, Chen Y, Xiao ZS, Chen X, Bu XL, Liu YH, Zeng F, Zhang L, Li A, Cai Y, Sun P, He Z, Doré V, Fripp J, Bourgeat P, Chen Q, Yu JT, Tang Y, Zetterberg H, Masters CL, Guo T, Wang YJ; Translational Biomarker Research of AgIng and Neurodegeneration (TBRAIN). Diagnostic accuracy of plasma p-tau217/A $\beta$ 42 for Alzheimer's disease in clinical and community cohorts. *Alzheimers Dement.* 2025 Mar;21(3):e70038. doi: 10.1002/alz.70038.