SUPPLEMENTAL APPENDIX

RELIABILITY OF MULTI-CENTER FDG-PET ASSESSMENTS

METHODS

Previous multi-center studies performed using similar image averaging and pre-processing procedures have shown that differences in spatially normalized FDG-PET scans obtained with scanners of different resolutions can be minimized by: (1) restricting the analysis to voxels with intensity 80% greater than the whole-brain (WB) mean, and (2) excluding voxels from the uppermost 10 slices (i.e., from the top 22.5 mm of the brain) and from the lowermost 5 slices, where significant inter-scanner effects due to different FOV were previously reported[28,39]. In the present study, we adopted the above precautions and further restricted the analysis to voxels within predefined ROIs, as listed above. CMRglc means were computed for each ROI and the CMRglc values averaged across hemispheres. Since HIP CMRglc is substantially lower than the whole-brain mean, HIP ROI values are expressed as HIP/whole-brain * 2. For examination of variability in voxel intensities across scanners, CMRglc measures were first normalized to the mean voxel intensity per image. The normalized CMRglc ROI measures were used to examine the feasibility of combining data acquired with different scanners by assessing whether: (a) multi-center variance in CMRglc measures leads to lower effect sizes in detecting disease, and (b) it is feasible to combine multi-center data to develop standardized and consistent FDG-PET diagnostic criteria, as follows:

1) Multi-center reliability in CMRglc measures.

To account for differences in sample sizes across centers, FDG-PET scans were randomly selected to include at least 5 subjects per diagnostic group per center. A total of 190 subjects were examined, including 50 NL, 50 MCI, 50 AD, 25 FTD and 15 DLB subjects balanced for age, gender, and education (Supplemental Appendix Table). The General Linear Model (GLM)/univariate analysis with post-hoc LSD tests and Chi square (χ^2) tests were used to examine clinical and CMRglc measures by testing for group effects, center effects and group by center effects on ROI CMRglc. First, coefficients of variation ([%] CV = SD/mean * 100) in each ROI across and within centers were examined. CVs $\leq$ 20% are considered acceptable. Second, we examined whether a multi-center approach would reduce sensitivity in detecting CMRglc reductions across clinical groups in comparison to a mono-center analysis. Statistical analyses were computed for each center using a two-tailed independent sample t-test of the group comparison and calculating the corresponding effect size (ES) of each ROI in distinguishing disease from NL aging ("mono-center ES"). The ES *d* of the difference in CMRglc between NL and each clinical group is estimated as: $d = (X-Y)/SD_{xy}$, where X-Y is the difference in the mean CMRglc between groups and SD_{xy} is the pooled standard deviation of CMRglc obtained from the two groups. ES with $d > 1$ are considered biologically meaningful. The ES of group differences across centers was then calculated for each ROI by using the mean CMRglc and pooled SD across centers ("multi-center ES"), and the multi-center ES compared to the mono-center ES.

2) Multi-center reliability in detecting CMRglc abnormalities using voxel-based analysis and Z scores.

FDG-PET scans of each subject were compared and Z-scored relative to the NL database using Neurostat[30,31] and HipMask[13]. CMRglc Z scores≥2.33 within the brain regions showing significant group differences in the above analysis were considered indicative of significant hypometabolism, and each FDG-PET scan was classified as Positive or Negative for CMRglc abnormalities based on the Z scores. Pearson's *r* correlation coefficients ($P<0.05$, one-tailed) were used to examine the relationship between Z scores derived from cortical ROIs and Z scores extracted from 3D-SSP maps for each scan. χ^2 tests were used to test for center effects and center by clinical group effects in detecting CMRglc abnormalities. Results were considered significant at $P<0.05$.

RESULTS

1) Multi-center reliability in CMRglc measures.

ROI CMRglc data are presented in what follows. After global normalization, CVs were comparable across centers, ROIs, and clinical groups (Supplemental Appendix Table). Across centers, CVs ranged from 3% for most scanners to a maximum of 19% for scanner II. Variability across CVs was mostly dependent on diagnosis and ROIs rather than scanner type. Across clinical groups, CVs ranged from 3% to 10% across scanners and ROIs in NL, from 3% to 18% in MCI, from 3% to 18% AD patients, from 2% to 10% in DLB, and from 2% to 11% in FTD. Across ROIs, CVs were generally larger in clinical groups as compared to NL. Specifically, CVs were larger in the HIP of MCI (10-18%), AD (11-19%) and FTD (4-11%) patients as compared to NL (3-8%), and in the

IPL and PCC of MCI (5-14%, and 5-15%, respectively) and AD patients (3-18%, and 4-15%) as compared to NL (3-8%, and 3-10%). Within-scanner CMRglc group effects are reported in the Supplemental Appendix Table.

Examination of the pooled multi-center CMRglc measures showed CVs <2 0%, which ranged from 7% (S-M) to 16% (HIP) in AD, 4% (S-M) to 11% (IPL) in DLB, 3% (S-M) to 10% (LTL) in FTD, from 5% (S-M) to 16% (HIP) in MCI, and from 4% (S-M) to 8% (PFC) in NL. Comparison of the pooled multi-center CMRglc measures to the NL database showed significant CMRglc group effects for HIP, PCC, IPL, LTL, PFC and OCC (*P's*$\leq$0.001). No group effects were found for the S-M CMRglc. On post-hoc analysis, there were no CMRglc differences between the NL subjects included in the database and the remaining NL in any ROIs.

AD patients showed significantly reduced CMRglc in the HIP as compared to NL (25%), DLB (24%), FTD (13%) and MCI (8%), in the IPL as compared to NL (18%), MCI (10%), and FTD (7%), and in the PCC as compared to NL (17%), DLB (10%) and FTD (8%) (*P's*$\leq$0.01). AD patients also showed reduced LTL and PFC CMRglc as compared to NL (8% and 9%, *P's*<0.001).

DLB patients showed reduced CMRglc in the OCC as compared to all other groups (16-18%), in the IPL as compared to NL (23%), MCI (16%), and FTD (13%), and in the PCC as compared to NL (10%) (*P's*$\leq$0.01).

FTD patients showed reduced CMRglc in the PFC as compared to NL (12%), MCI and DLB (8%), in the LTL as compared to NL (13%) and MCI (10%), and in the HIP as compared to NL (14%) (*P's*$\leq$0.03).

MCI patients showed reduced CMRglc in the HIP as compared to NL (19%) and DLB (18%), and in the PCC as compared to NL (12%) (P's$\leq$0.001), and showed a trend towards reduced IPL CMRglc as compared to NL (9%, P=0.09).

Overall, in comparison to the database of NL subjects, data from all centers showed that: (i) PCC, HIP, IPL and LTL yielded the largest ES in distinguishing AD from NL, (ii) OCC and IPL yielded the largest ES in distinguishing DLB from NL, (iii) PFC and LTL, and to lesser extent HIP, yielded the largest ES in distinguishing FTD from NL, and (iv) PCC and HIP yielded the largest ES in distinguishing MCI from NL.

Within these ROIs, mono-center CMRglc data showed generally large ES in separating NL from clinical groups. In the comparison between AD and NL, the ES ranged from 2.1-4.3 for the PCC, from 1.8-3.3 for the HIP, from 1.8-3.7 for the IPL, and from 1.2-3.0 for the LTL. In the comparison between DLB and NL, the ES ranged from 2.1-3.2 for the OCC, and from 1.0-2.5 for the IPL. In the comparison between FTD and NL, the ES ranged from 2.9-3.4 for the PFC, from 1.7-2.6 for the LTL, and from 1.0-1.9 for the HIP. In the comparison between MCI and NL, the ES ranged from 1.5-4.0 for the PCC, and from 1.2-2.4 for the HIP.

The pooled multi-center CMRglc data also yielded large ES in all affected ROIs. The multi-center ES for the comparison between AD and NL were 3.2 for the PCC, 2.8 for the HIP, 2.1 for the IPL, and 2.0 for the LTL. The ES for the comparison between DLB and NL were 2.8 for the OCC and 1.6 for the IPL. The ES for the comparison between FTD and NL were 3.2 for the PFC, 1.9 for the LTL, and 1.5 for the HIP. The ES for the comparison between MCI and NL was 3.0 for the PCC and 2.0 for the HIP.

These results show that multi-center variability at the voxel level can be accounted for by focusing on ROIs showing consistent CMRglc abnormalities with large ES on a mono-center basis.

2) Multi-center reliability in detecting CMRglc abnormalities using voxel-based analysis and Z scores.

Z scores extracted from 3D-SSP maps are examined in what follows. ROI CMRglc data are presented in what follows. FDG-PET scans of every subject were Z-scored relative to the NL database, and scans classified as Positive or Negative for the presence of a neurodegenerative disease based on the estimated Z scores within the affected ROIs described above. Significant correlations were found between Z scores derived from the ROIs and Z scores extracted from 3D-SSP maps. For the entire group, the correlation coefficients were r=0.79 for IPL, r=0.75 for LTL, r=0.82 for PCC, r=0.90 for PFC, r=0.87 for OCC and r=0.92 for S-M (*P's*<0.001). Within each clinical group, the correlation coefficients ranged from r=0.77 (LTL) to r=0.94 (S-M) in NL, from r=0.77 (PFC) to r=0.92 for PCC in AD, from r=0.86 (OCC) to r=0.98 (PFC and PCC) in FTD, from r=0.74 (LTL) to r=0.89 (OCC) in DLB, and from r=0.73 (LTL) to r=0.89 (PCC) in MCI (*P's*$\leq$0.001).

A Positive PET pattern was found in 92% AD, 90% DLB, 85% FTD, and 80% MCI. A Negative PET pattern was found in 89% NL. No differences in the number of cases correctly identified across scanners, and no scanner by clinical group effects were found (*P's*>0.1, n.s.).