Visual-Statistical Interpretation of ^{18}F-FDG-PET Images for Characteristic Alzheimer Patterns in a Multicenter Study: Inter-Rater Concordance and Relationship to Automated Quantitative Evaluation

T. Yamane, Y. Ikari, T. Nishio, K. Ishii, K. Ishii, T. Kato, K. Ito, D.H.S. Silverman, M. Senda, T. Asada, H. Arai, M. Sugishita, T. Iwatsubo, and the J-ADNI Study Group

ABSTRACT

BACKGROUND AND PURPOSE: The role of ^{18}F-FDG-PET in the diagnosis of Alzheimer disease is increasing and should be validated. The aim of this study was to assess the inter-rater variability in the interpretation of ^{18}F-FDG-PET images obtained in the Japanese Alzheimer's Disease Neuroimaging Initiative, a multicenter clinical research project.

MATERIALS AND METHODS: This study analyzed 274 ^{18}F-FDG-PET scans (67 mild Alzheimer disease, 100 mild cognitive impairment, and 107 normal cognitive) as baseline scans for the Japanese Alzheimer's Disease Neuroimaging Initiative, which were acquired with various types of PET or PET/CT scanners in 23 facilities. Three independent raters interpreted all PET images by using a combined visual-statistical method. The images were classified into 7 (FDG-7) patterns by the criteria of Silverman et al and further into 2 (FDG-2) patterns.

RESULTS: Agreement among the 7 visual-statistical categories by at least 2 of the 3 readers occurred in >94% of cases for all groups: Alzheimer disease, mild cognitive impairment, and normal cognitive. Perfect matches by all 3 raters were observed for 62% of the cases by FDG-7 and 76 by FDG-2. Inter-rater concordance was moderate by FDG-7 ($\kappa = 0.57$) and substantial in FDG-2 ($\kappa = 0.67$) on average. The FDG-PET score, an automated quantitative index developed by Herholz et al, increased as the number of raters who voted for the AD pattern increased ($\rho = 0.59$, $P < .0001$), and the FDG-PET score decreased as those for normal pattern increased ($\rho = -0.64$, $P < .0001$).

CONCLUSIONS: Inter-rater agreement was moderate to substantial for the combined visual-statistical interpretation of ^{18}F-FDG-PET and was also significantly associated with automated quantitative assessment.

ABBREVIATIONS: AD = Alzheimer disease; J-ADNI = Japanese Alzheimer's Disease Neuroimaging Initiative; MCI = mild cognitive impairment; NC = cognitively normal subject

PET can visualize regional glucose metabolism by using ^{18}F-FDG; and hypometabolism in the posterior cingulate/precuneus and temporoparietal cortices is regarded as a typical uptake pattern of Alzheimer disease (AD).[1] These findings are considered useful for differentiating AD from other disorders presenting with dementia as well as for predicting conversion from mild cognitive impairment (MCI) to AD.[2,3]

Three approaches for evaluating brain PET images are visual interpretation alone, visual interpretation with adjunctive statis-

Received February 21, 2013; accepted after revision May 2.

From the Division of Molecular Imaging (T.Y., Y.I., T.N., M. Senda), Institute of Biomedical Research and Innovation, Kobe, Japan; Department of Radiology (Kazunari Ishii), Kinki University Faculty of Medicine, Osakasayama, Japan; Positron Medical Center (Kenji Ishii), Tokyo Metropolitan Institute of Gerontology, Tokyo, Japan; Department of Brain Science and Molecular Imaging (T.K., K. Ito), National Center for Geriatrics and Gerontology, Obu, Japan; David Geffen School of Medicine (D.H.S.S.), University of California, Los Angeles, Los Angeles, California; Department of Psychiatry (T.A.), University of Tsukuba, Tsukuba, Japan; Department of Geriatrics and Gerontology (H.A.), Tohoku University, Sendai, Japan; Institute of Brain and Blood Vessels (M. Sugishita), Isezaki, Japan; Department of Neuropathology and Neuroscience (T.I.), University of Tokyo, Tokyo, Japan; Research Association for Biotechnology (Y.I., T.N.), Tokyo, Japan; and J-ADNI Core (T.Y., Y.I., T.N., Kazunari Ishii, Kenji Ishii, T.K., Kengo Ito, M.S., T.A., H.A., M.S., T.I.).

The Research Group of the Japanese Alzheimer's Disease Neuroimaging Initiative comprised investigators from 38 different facilities. The investigators contributed to the design and implementation of J-ADNI and/or provided data but did not participate in the analyses of this report.

T. Yamane contributed to concept and design, analyzed data, and wrote the manuscript. Y. Ikari and T. Nishio acquired and analyzed PET data. Kazunari Ishii, Kenji Ishii, T. Kato, and K. Ito acquired and interpreted PET data. D.H.S. Silverman critically revised the manuscript and enhanced its intellectual content. M. Senda critically revised the manuscript, enhanced its intellectual content, and approved the final content of the manuscript. T. Asada, H. Arai, M. Sugishita, and T. Iwatsubo acquired clinical data and approved the final content of the manuscript.

This work is a part of the Translational Research Promotion Project/Research Project for the Development of a Systematic Method for the Assessment of Alzheimer's Disease, sponsored by the New Energy and Industrial Technology Development Organization of Japan. The Japanese Alzheimer's Disease Neuroimaging Initiative is also supported by a Grant-in-Aid for Comprehensive Research on Dementia from the Japanese Ministry of Health, Labour and Welfare, as well as by the grants from J-ADNI Pharmaceutical Industry Scientific Advisory Board companies.

Paper previously presented in part at: Annual Meeting of the Society of Nuclear Medicine, June 4–8, 2011; San Antonio, Texas.

Please address correspondence to Tomohiko Yamane, MD, PhD, Division of Molecular Imaging, Institute of Biomedical Research and Innovation, 2–2, Minatojima-minamimachi, Chuo-ku, Kobe, 650-0047, Japan; e-mail address: yamane@fbri.org

Indicates open access to non-subscribers at www.ajnr.org

Evidence-Based Medicine Level 2.

http://dx.doi.org/10.3174/ajnr.A3665

tical tools (visual-statistical), and automated quantitative analysis, but the relationship between the latter 2 of these approaches has been little explored, to our knowledge. Visual interpretation features comprehensive and flexible assessment of the qualitative radioactivity distribution by the reader, who may look into all features across the brain. This approach appears effective because patients with AD typically present with characteristic temporoparietal hypometabolism known as the "AD pattern." However, inter-rater variability inevitably occurs because each rater has his or her own experience and criteria, especially for borderline cases, and this variability can potentially be increased or decreased when the reader also takes into account statistical information provided by various software display tools.

On the other hand, quantitative analysis traditionally extracts radioactivity uptake values of the region of interest, placement of which is a subjective matter requiring experience. Although a recently developed anatomic standardization technique can define ROIs automatically and further allows voxelwise statistical analysis to generate *Z*-maps, standardization may not always be accurate and may require adjustment by a human observer. Although these region-of-interest values can be processed into a numeric indicator such as an FDG-PET score[4,5] and a cutoff level can be determined, a single indicator may not be as accurate as complex and comprehensive evaluation by expert readers. As a result, a "combined" approach of visual and quantitative evaluation is often used during image interpretation, in which the readers examine both the tomographic PET images and the result of region-of-interest analysis and/or a *Z*-map.

Inter-rater variability and comparison between visual reading and software-based evaluation have been studied by some investigators on brain ^{18}F-FDG-PET. Ng et al[6] studied the inter-rater variability of 15 patients with AD and 25 cognitively normal subjects (NCs) and reported that visual agreement between 2 readers was good ($\kappa = 0.56$). Tolboom et al[7] studied the variability of 20 patients with AD and 20 NCs and reported that agreement between 2 readers was moderate ($\kappa = 0.56$). Rabinovici et al[8] also reported the inter-rater agreement of ^{18}F-FDG ($\kappa = 0.72$). However, the data of these preceding studies were acquired with a single scanner in a single site and were evaluated by the readers belonging to the institution who were used to the scanner and its image quality. In addition, the studied subjects did not include patients with MCI, in whom PET findings featuring AD, if any, are mild and may make the discrimination challenging. Furthermore, inter-rater variability for combined interpretation of visual and statistical analysis has never been reported, to our knowledge.

In the present study, we analyzed the baseline scans of ^{18}F-FDG in a multicenter clinical project named Japanese Alzheimer's Disease Neuroimaging Initiative (J-ADNI)[9] and evaluated the inter-rater variability among 3 independent expert raters who were blinded to the clinical information and interpreted the PET images to evaluate the characteristic AD pattern in ^{18}F-FDG-PET on the basis of a combined visual-statistical evaluation. The raters looked at the 3D stereotactic surface projection *Z*-map of ^{18}F-FDG-PET visually as well as the ^{18}F-FDG tomographic images because it is considered the standard means of human interpretation of ^{18}F-FDG-PET images in Japan and therefore was adopted as the official interpretation method in J-ADNI. Images were also assessed by automated quantitative analysis by using an FDG-PET score, which was derived from ADtsum,[4,5] and were compared with the visual-statistical rating by the 3 raters and with their consensus.

MATERIALS AND METHODS

Subjects

Data used in the present study were obtained from J-ADNI.[9] This project was approved by the ethics committee of each site in which J-ADNI data were acquired, and written informed consent was obtained from each subject before participating in J-ADNI. All subjects were native Japanese speakers, 60–84 years of age, and were registered as 1 of 3 clinical groups (mild AD, MCI, or NC). Subjects of the mild AD group scored 20–26 in Mini-Mental State Examination-Japanese and 0.5–1.0 in the Clinical Dementia Rating-Japanese and were compatible with the probable AD criteria in the National Institute of Neurologic and Communicative Disorders and Stroke and the Alzheimer Disease and Related Disorders Association.[10] Subjects of the MCI group scored 24–30 in the Mini-Mental State Examination-Japanese and 0.5 in the Clinical Dementia Rating-Japanese. Subjects of NC group scored 24–30 in the Mini-Mental State Examination-Japanese and 0 in the Clinical Dementia Rating-Japanese. The exclusion criteria were depression (Geriatric Depression Scale-Japan $\geq$ 6), cerebrovascular disorders (Hachinski Ischemic Score $\geq$ 5), and other neurologic or psychiatric disorders.

Enrollment in each clinical group for J-ADNI was primarily determined by the referring physician, and 303 consecutive subjects entered the study to undergo ^{18}F-FDG-PET scanning. A thorough central review of the clinical and behavioral data by expert psychiatrists and psychologists excluded 29 cases that had erroneous assessment of the cognitive test results, depression or cerebrovascular disorders that had been overlooked, prohibited concomitant medications, or other deviations from the criteria. As a result, 274 baseline ^{18}F-FDG-PET scans (67 mild AD, 100 MCI, and 107 NC) were analyzed in the present study.

PET Imaging

As a quality assurance measure necessary for the multicenter study, all PET sites in J-ADNI were qualified for the PET scanner and other devices, resting-state environment, quality of the on-site-produced PET drugs, and so forth before scanning of the first subject. Intersite differences were minimized by standardizing the imaging protocol, and interscanner differences were addressed with the Hoffmann 3D phantom data.[11] The data used for the analysis in the present study were acquired with 14 types of PET or PET/CT scanners in 23 PET centers.

In the ^{18}F-FDG-PET scans, all subjects fasted for at least 4 hours and their preinjection blood glucose levels were confirmed to be <180 mg/dL. Intravenous administration of ^{18}F-FDG (185 $\pm$ 37 MBq) was followed by a resting period of 30 minutes in a dimly lit and quiet room. Dynamic scans (300 seconds $\times$ 6 frames) were obtained starting 30 minutes postinjection in the 3D mode. Attenuation was corrected for by a transmission scan with segmentation for dedicated PET and by a CT scan for PET/CT.

All the PET images acquired in each PET site went through the J-ADNI PET quality control process,[11] in which head motion between frames was corrected for and bad frames were removed to create sum frame images. Then the images were reoriented to the anterior/posterior commissure line with the same matrix size and

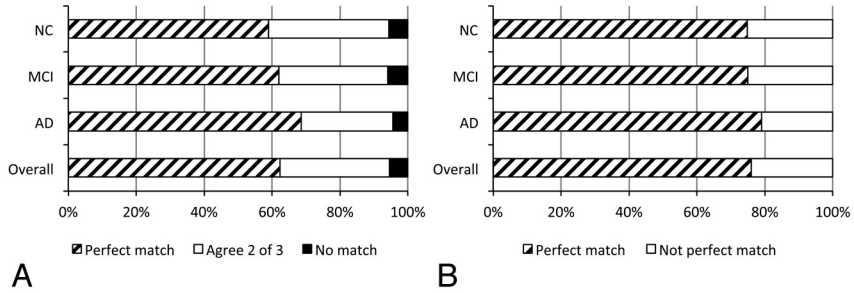

FIG 1. Breakdown of the ^{18}F-FDG-PET cases into degree of match by 3 raters in a combined visual-statistical human classification into 7 (FDG-7) (*A*) or 2 (FDG-2) (*B*) categories. A perfect match by the 3 raters is observed for 62% of the cases for FDG-7 and 76% for FDG-2 in total. The AD group shows the highest concordance followed by the MCI and NC groups, in this order, both for FDG-7 and FDG-2.

voxel size so that all camera models presented images of similar orientation and appearance to the viewer and were then passed on to image interpretation.

The ^{18}F-FDG-PET images that had passed through the quality control process above were also treated with a 3D stereotactic surface projection technique to generate *z* score maps (displayed with upper = 7 and lower = 0) by using iSSP software, Version 3.5 (Nihon Medi-physics, Tokyo, Japan). The normal data base used for generating the *Z*-maps was made by a method of leave-one-out cross-validation based on 25 healthy subjects of J-ADNI (11 men and 14 women; mean age, 66.0 ± 4.8 years) who were interpreted as having a normal pattern by one of the coauthors of the study. The *Z*-maps were used not for the automated quantification but for a part of the information for human raters in the visual-statistical interpretation.

Human Interpretation

Those ^{18}F-FDG images generated through the quality control process above were independently interpreted with the combined visual-statistical method by 3 expert raters blinded to the clinical group and other clinical and laboratory data. The raters were provided with the ^{18}F-FDG tomographic images on the viewer as well as the *Z*-map images in PDF format. Information about the age and sex was also provided to the raters. Moreover, T1-weighted MR images acquired in 3D mode by using MPRAGE or its equivalent and reformatted in axial sections were also provided together with axial T2WI and proton-attenuation images, in which the MR imaging sections did not correspond to the PET section positions. The experience of the 3 raters as physicians specializing in nuclear neuroimaging was 17, 19, and 19 years, respectively, when this project started.

After independent interpretation, consensus reads were performed by the 3 raters and 2 other discussants who are experienced nuclear medicine physicians specialized in neuroimaging. The experience of both discussants as physicians specializing in nuclear neuroimaging was 20 years. The same images and information as that in the independent interpretation were also provided for the discussants in the consensus reads. The 7 sessions of consensus reads lasted for 1.5 years in the order of subject enrollment in J-ADNI. In the consensus reads, the cases in which the evaluations by the 3 raters did not completely match were discussed, and the unified visual-statistical interpretation was determined as an official judgment by the J-ADNI PET Core.

For classification of ^{18}F-FDG-PET, the criteria of Silverman et al[1] were adopted for classifying the uptake pattern in J-ADNI. All 3 expert raters and the 2 discussants had attended a training course for the criteria organized by Silverman et al before starting the J-ADNI project. In the criteria of Silverman et al, ^{18}F-FDG uptake patterns were classified into 7 categories: progressive patterns: P1, P1+, P2, and P3, in which P1 represents the characteristic AD pattern and P1+ represents AD-variant pattern, including the characteristic Lewy body dementia pattern; and nonprogressive patterns: N1, N2 and N3, in which N1 represents the characteristic normal pattern. In addition to these original 7 categories (FDG-7), the present study defined a binary criteria (FDG-2) in which the 7 categories were dichotomized into posterior-predominant hypometabolism (AD and AD-variant) patterns (P1, P1+) and the other patterns (N1, N2, N3, P2, and P3).

Automated Quantitative Evaluation

In the automated quantitative analysis, the FDG-PET score, as a measure of the AD pattern, was calculated from ADtsum[4] by using the Alzheimer's Discrimination Tool in PMOD, Version 3.12 (PMOD Technologies, Zurich, Switzerland)[4,5] by using the following equation: FDG-PET score = log2 {(ADtsum / 11,089) + 1}. The FDG-PET score was not calculated in 1 case because no significant clusters were determined for the image.[4] This case was excluded from the quantitative analysis.

Statistical Analysis

Concordance among the 3 raters was evaluated by Cohen κ statistics. As comparisons between human and automated evaluation, the association between the FDG-PET score and the number of the raters who interpreted the case as P1 (AD pattern) in FDG-7 was evaluated by the Spearman rank correlation coefficient. Likewise, association between the FDG-PET score and the number of the raters who interpreted the case as N1 (normal pattern) was evaluated. The association was also examined between the FDG-PET score and the number of raters in FDG-2 classification (ie, how many raters judged the case as the AD and AD-variant patterns [P1, P1+] versus the other patterns [N1, N2, N3, P2, and P3]). A *P* value < .05 was considered significant. In addition, the FDG-PET score was compared with the final combined visual-statistical interpretation determined by the consensus read and with the clinical group. Receiver operating characteristic analysis was used to obtain the optimum cutoff level for the quantitative index for discrimination.

Neither iSSP nor the PMOD Alzheimer's Discrimination Tool was approved for clinical use by the US Food and Drug Administration.

RESULTS

Figure 1 summarizes concordance rates among the 3 raters. Agreement among the 7 visual-statistical categories by at least 2 of the 3 readers occurred in >94% of cases for all groups: NC, MCI,

and AD. The κ statistic $\pm$ SE for each pair of the 3 raters was 0.59 $\pm$ 0.04, 0.54 $\pm$ 0.04, and 0.58 $\pm$ 0.04 in FDG-7 (average, 0.57), and 0.73 $\pm$ 0.04, 0.65 $\pm$ 0.0, and 0.64 $\pm$ 0.05 in FDG-2 (average, 0.67), respectively.

Figure 2 illustrates the relationship between the FDG-PET score and the number of raters who visually-statistically interpreted the ^{18}F-FDG-PET image as P1 (Fig 2*A*) and N1 (Fig 2*B*). A significant positive association was observed between the FDG-PET score and the number of P1 interpretations ($\rho = 0.59$, $P < .0001$). The mean FDG-PET score was 0.46 $\pm$ 0.37 ($n = 103$) for the scans no raters interpreted as P1, but it increased to 0.723 $\pm$ 0.39 ($n = 34$) for those that 1 rater interpreted as P1, to 0.99 $\pm$ 0.45 ($n = 31$) for 2 raters, and to 1.21 $\pm$ 0.73 ($n = 105$) for all 3 raters. Likewise, a significant negative association was observed between the FDG-PET score and the number of N1 interpretations ($\rho = -.64$, $P < .0001$). The FDG-PET score was 1.15 $\pm$ 0.69 ($n = 146$) for the scans no raters interpreted as N1, but it decreased to 0.80 $\pm$ 0.39 ($n = 28$) for those 1 rater interpreted as N1, 0.50 $\pm$ 0.25 ($n = 40$) for 2 raters, and 0.34 $\pm$ 0.22 ($n = 59$) for all 3 raters. A similar association was observed between the FDG-PET score and the number of raters who interpreted the case as AD and AD-variant patterns, including the Lewy body dementia pattern (P1, P1+) or the other patterns (N1, N2, N3, P2, and P3); and both showed significant positive and negative associations ($\rho = 0.60$, $P < .0001$; and $\rho = -0.60$, $P < .0001$).

Figure 3 illustrates scatterplots of the FDG-PET scores as contrasted to the combined visual-statistical interpretation determined by the consensus read of ^{18}F-FDG-PET for each clinical group. For each group as well as for all subjects, cases with P1 interpretation showed higher FDG-PET scores than those with N1. Receiver operating characteristic analysis on P1 and N1 cases led to a cutoff FDG-PET score of 0.67 for discrimination between P1 and N1. As was expected, NC cases with P1 interpretation had lower FDG-PET scores than MCI and AD cases with P1 interpretation, and the ratio of the cases above-to-below the cutoff level was also lower. As for the cases with other patterns, a large fraction of the cases with N2 interpretation had FDG-PET scores above the cutoff level, though most were below 1.0. The FDG-PET scores of the cases with P1+ and P2 were variable.

DISCUSSION

Matches among 7 visual-statistical categories by at least 2 of 3 readers occurred in >94% of cases for each clinical group, and perfect matches among the 3 raters were observed for 62% of the cases for FDG-7 and 76% for FDG-2 categorization schemes in total. The mild AD group showed the highest concordance, followed by MCI and NC, in order, for both FDG-7 and FDG-2. The AD pattern in ^{18}F-FDG-PET is usually seen in the early stage of AD and is expected to predict the onset of AD.[1,12] Because most of the subjects who are clinically diagnosed as having AD may have had an established AD pattern in ^{18}F-FDG-PET, it is reasonable for these results that AD showed the highest concordance.

Based on the classification of κ values described by Landis and Koch,[13] agreements were considered to be moderate for FDG-7 and substantial for FDG-2. Inter-rater variability is one of the indices that are often used to evaluate the validity of methods of image interpretation, and it facilitates comparison with the other studies. The κ index of FDG-2 ($\kappa = 0.67$) of the present study showed values similar to those of the other studies ($\kappa = 0.56$-$.72$) evaluated by the bi-

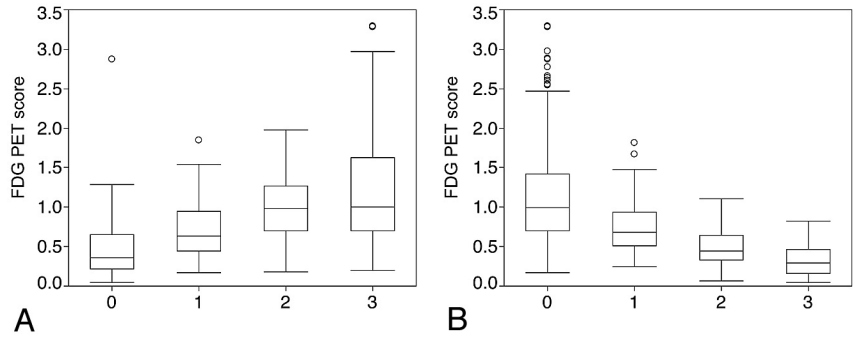

FIG 2. Boxplots of the FDG-PET score against the number of raters who interpreted the ^{18}F-FDG-PET images as P1 (*A*) and N1 (*B*) based on the FDG-7 criteria. The FDG-PET score gradually increases as the number of P1 (AD pattern) interpretations increases (Spearman rank correlation coefficient: $\rho = 0.59$, $P < .0001$). On the other hand, FDG-PET score gradually decreases as the number of N1 (normal pattern) interpretations increases ($\rho = -.64$, $P < .0001$).

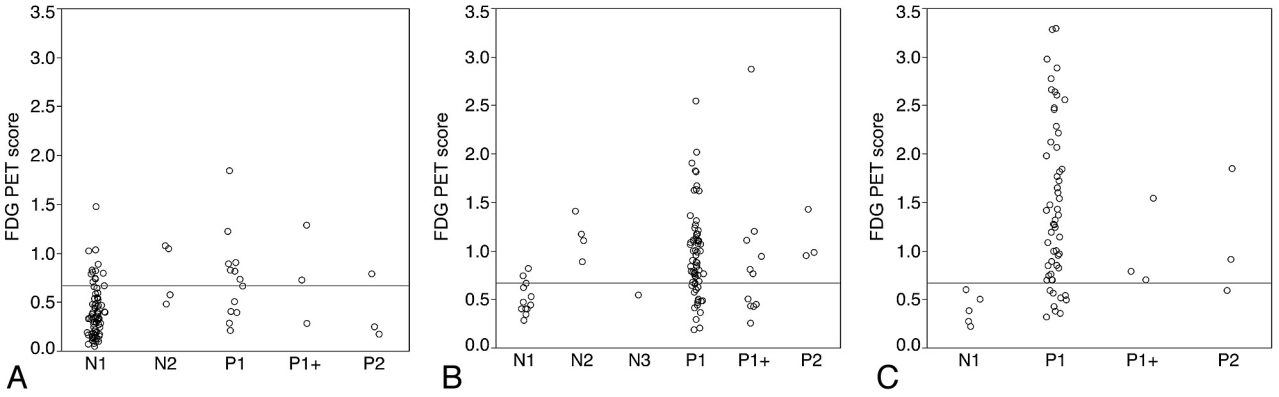

FIG 3. Scatterplot of the FDG-PET score as contrasted with the combined visual-statistical interpretation determined by the consensus read of ^{18}F-FDG-PET for each clinical group (*A*, NC; *B*, MCI; and *C*, AD). The horizontal line indicates the cutoff level of 0.67 derived by receiver operating characteristic analysis on P1 and N1 cases.

nary criteria.[6-8] However, the values observed in the other studies are not the same as those in the present study because we analyzed the interpretation both visually and statistically. Recent studies have shown that the diagnostic capability of visual analysis of ^{18}F-FDG-PET increases when the raters interpret the images in combination with 3D stereotactic surface projections.[14,15] These kinds of visual-statistical methods seem to be a standard approach in clinical settings.

To increase the concordance rate and diagnostic capability, we need to overcome some problems. We had to degrade the image quality according to the PET with the lowest quality among the 23 facilities of J-ADNI.[11] Therefore, the quality of the images may be improved in the future. In addition to the image quality, development of new methods or new approaches to image interpretation may contribute to increasing the concordance.

This study showed a relationship between combined visual-statistical interpretation and automated quantitative assessment regarding the characteristic AD pattern in brain ^{18}F-FDG-PET. Significant association was observed between the quantitative index (FDG-PET score) and the number of raters who interpreted the scans accordingly. This correlation may have been something expected from reports on similar/automated analysis.[5,6] However, this association was observed in a large-scale multicenter study by using various camera models on a wide spectrum of subjects in the present study.

From the standpoint of detecting the AD pattern, cases evaluated as having positive AD findings by complete agreement of all 3 raters tended to show a higher quantitative index than the cases that fewer than 3 raters interpreted as having positive AD findings. From the standpoint of ruling out the AD pattern, cases evaluated as having negative AD findings by complete agreement of all 3 raters also tended to show a lower quantitative index than the cases that fewer than 3 raters interpreted as having negative AD findings. Therefore, the results suggest that interpretation by 3 raters may be better than that by 2 or fewer raters. The results also indicate that cases that only 1 rater interpreted as having positive (or negative) AD findings presented a different quantitative index from those that no raters interpreted as having positive (or negative) findings. This outcome suggests that there are cases in which the "minority opinion" may not be ignored.

Generally, the minority opinion is somewhat important when a subtle but definite finding is evaluated. However, most of the ^{18}F-FDG-PET images for which the judgment did not agree among the raters showed ambiguous findings. Ng et al[6] reported that experienced raters scored higher accuracy than nonexperienced raters in the interpretation of brain ^{18}F-FDG-PET images for the diagnosis of AD.[6] Such subtle findings in brain ^{18}F-FDG-PET may be difficult to interpret. We need to analyze the difference in detail and develop new methods for interpretation or new diagnostic tools.

When the FDG-PET score of the cases judged as P1 in the consensus read were examined, NC subjects with P1 interpretation showed lower FDG-PET scores than MCI and AD subjects. This result is probably because many of the NC subjects with P1 interpretation presented with a very mild AD pattern that influenced the FDG-PET score to only a small extent. Those cases, however, presented characteristic findings such as posterior cingulate hypometabolism, which led to the P1 interpretation.

The criterion standard used in this study was the clinical diagnosis at enrollment. Although dementia with Lewy body cases with the specific symptoms were excluded from enrollment in the J-ADNI beforehand, differentiating Lewy body dementia from AD is occasionally difficult in clinical settings.[16] The typical Lewy body dementia pattern of ^{18}F-FDG-PET, evaluated as occipital hypometabolism, is classified into P1+ by the criteria of Silverman et al.[1] Some cases classified into P1+, though limited in the present study, seem to have the possibility of Lewy body dementia. Moreover, the consensus read judged 16 of 107 cases of the NC group to be the AD pattern (P1 and P1+), and 8 of 67 cases in the AD group to be a non-AD pattern (N1 and P2). These disagreements might be either caused by inappropriate clinical diagnosis at enrollment or reflecting the limitation of FDG-PET as a diagnostic tool. While these diagnostic discrepancies are not critical in the present study, which analyzed inter-rater concordance, comparison with other criterion standards such as long-term follow-up or postmortem examination is important for this kind of multicenter study in the future.

The FDG-PET score of 1.0, by definition, is proposed as an optimum threshold for the differential diagnosis of AD from healthy subjects.[5] Because the present study deals with comparison of combined visual-statistical human interpretation with automated quantitative analysis, we derived a cutoff level of 0.67 based on discrimination of the P1 from the N1 pattern. This discrepancy may be explained by the difference in the target of discrimination as well as in the profile of subjects, and the lower cutoff would be consistent with a higher sensitivity for visually detecting the AD pattern than for clinically identifying the diagnosis of AD, for which the 1.0 cutoff is designed. In addition, one of the essential factors for this discrepancy seems to be that decisions by visual-statistical interpretation are not completely consistent with the actual clinical diagnosis. Because the diagnostic capability of ^{18}F-FDG-PET is not the subject of the present study, further studies are needed to elucidate the discrepancy.

CONCLUSIONS

Inter-rater agreement was moderate to substantial regarding the combined visual-statistical human interpretation of the characteristic AD pattern in ^{18}F-FDG-PET. In addition, a significant relationship between human interpretation and automated quantitative assessment was found. The human rating as an AD or normal pattern was best predicted by the FDG-PET score when using a cutoff of 0.67.

ACKNOWLEDGMENTS

The authors thank the J-ADNI Imaging Pharmaceutical Industry Scientific Advisory Board and other organizations for their support of this work.

Disclosures: Tomohiko Yamane—*RELATED*: *Grant*: New Energy and Industrial Technology Development Organization of Japan,* *Comments*: This study is a part of the "Translational Research Promotion Project/Research Project for the Development of a Systematic Method for the Assessment of Alzheimer's Disease," sponsored by the New Energy and Industrial Technology Development Organization of Japan. The Japanese Alzheimer's Disease Neuroimaging Initiative is also supported by a Grant-in-Aid for Comprehensive Research on Dementia from the Japanese Ministry of

Health, Labour and Welfare, as well as by grants from J-ADNI Pharmaceutical Industry Scientific Advisory Board companies. Yasuhiko Ikari—*UNRELATED*: *Employment*: Micron. Tomoyuki Nishio—*UNRELATED*: *Employment*: Micron. Takashi Kato—*RELATED*: *Grant*: government-based funding, *Comments*: Ministry of Health, Labour, and Welfare in Japan; Ministry of Education, Culture, Sports, Science & Technology in Japan. Kengo Ito—*RELATED*: *Grant*: "Translational Research Promotion Project/Research Project for the Development of a Systematic Method for the Assessment of Alzheimer's Disease," sponsored by the New Energy and Industrial Technology Development Organization of Japan, and a Grant-in-Aid for Comprehensive Research on Dementia from the Japanese Ministry of Health, and Labour and Welfare.* Michio Senda—*RELATED*: *Grant*: Research Association for Biotechnology,* *UNRELATED*: *Grants/Grants Pending*: GE Healthcare,* Eli Lilly.* Morihiro Sugishita—*RELATED*: *Grant*: New Energy and Industrial Technology Development Organization,* *Support for Travel to Meetings for the Study or Other Purposes*: New Energy and Industrial Technology Development Organization, *UNRELATED*: *Consultancy*: Eli Lilly Japan KK, *Employment*: Sugishita Co Ltd, *Royalties*: Nihon Bunka Kagakusha Co Ltd, *Stock/Stock Options*: Japan Air Line. *Money paid to the institution.

REFERENCES

1. Silverman DH, Small GW, Chang CY, et al. **Positron emission tomography in evaluation of dementia: regional brain metabolism and long-term outcome.** *JAMA* 2001;286:2120–27
2. Mosconi L, Perani D, Sorbi S, et al. **MCI conversion to dementia and the APOE genotype: a prediction study with FDG-PET.** *Neurology* 2004;63:2332–40
3. Yuan Y, Gu ZX, Wei WS. **Fluorodeoxyglucose-positron-emission tomography, single-photon emission tomography, and structural MR imaging for prediction of rapid conversion to Alzheimer disease in patients with mild cognitive impairment: a meta-analysis.** *AJNR Am J Neuroradiol* 2009;30:404–10
4. Herholz K, Salmon E, Perani D, et al. **Discrimination between Alzheimer dementia and controls by automated analysis of multicenter FDG-PET.** *Neuroimage* 2002;17:302–16
5. Herholz K, Westwood S, Haense C, et al. **Evaluation of a calibrated ^{18}F-FDG-PET score as a biomarker for progression in Alzheimer disease and mild cognitive impairment.** *J Nucl Med* 2011;52:1218–26
6. Ng S, Villemagne VL, Berlangieri S, et al. **Visual assessment versus quantitative assessment of ^{11}C-PIB PET and ^{18}F-FDG-PET for detection of Alzheimer's disease.** *J Nucl Med* 2007;48:547–52
7. Tolboom N, van der Flier WM, Boverhoff J, et al. **Molecular imaging in the diagnosis of Alzheimer's disease: visual assessment of [^{11}C]PIB and [^{18}F]FDDNP PET images.** *J Neurol Neurosurg Psychiatry* 2010;81:882–84
8. Rabinovici G, Rosen H, Alkalay A, et al. **Amyloid vs FDG-PET in the differential diagnosis of AD and FTLD.** *Neurology* 2011;77:2034–42
9. Iwatsubo T. **Japanese Alzheimer's Disease Neuroimaging Initiative: present status and future.** *Alzheimers Dement* 2010;6:297–99
10. McKhann G, Drachman D, Folstein M, et al. **Clinical diagnosis of Alzheimer's disease: report of the NINCDS-ADRDA Work Group under the auspices of Department of Health and Human Services Task Force on Alzheimer's Disease.** *Neurology* 1984;34:939–44
11. Ikari Y, Nishio T, Makishi Y, et al. **Head motion evaluation and correction for PET scans with ^{18}F-FDG in the Japanese Alzheimer's Disease Neuroimaging Initiative (J-ADNI) multi-center study.** *Ann Nucl Med* 2012;26:535–44
12. Silverman DH, Mosconi L, Ercoli L, et al. **Positron emission tomography scans obtained for the evaluation of cognitive dysfunction.** *Semin Nucl Med* 2008;38:251–61
13. Landis JR, Koch GG. **The measurement of observer agreement for categorical data.** *Biometrics* 1977;33:159–74
14. Lehman VT, Carter RE, Claassen DO, et al. **Visual assessment versus quantitative three-dimensional stereotactic surface projection fluorodeoxyglucose positron emission tomography for detection of mild cognitive impairment and Alzheimer disease.** *Clin Nucl Med* 2012;37:721–26
15. Foster NL, Heidebrink JL, Clark CM, et al. **FDG-PET improves accuracy in distinguishing frontotemporal dementia and Alzheimer's disease.** *Brain* 2007;130(pt 10):2616–35
16. Ishii K, Soma T, Kono AK, et al. **Comparison of regional brain volume and glucose metabolism between patients with mild dementia with Lewy bodies and those with mild Alzheimer's disease.** *J Nucl Med* 2007;48:704–11