Optimization of Automated Quantification of ^{123}I-IBZM Uptake in the Striatum Applied to Parkinsonism

Perry E. Radau, Rainer Linke, Piotr J. Slomka, and Klaus Tatsch

Department of Medical Biophysics, University of Western Ontario, London, Ontario; Department of Nuclear Medicine, London Health Sciences Centre, London, Ontario, Canada; and Department of Nuclear Medicine, Klinikum Grosshadern, University of Munich, Munich, Germany

Evaluation of therapies for parkinsonism by dopamine receptor SPECT requires a reproducible, optimized quantitation technique. This study presents a new, objective, automated technique for semiquantitative analysis of dopamine receptor density, as applied to the differential diagnosis of parkinsonism. **Methods:** Dopamine receptor density measured by ^{123}I-iodobenzamide (IBZM) SPECT was retrospectively analyzed in nonidiopathic parkinsonism (NIPS), in Parkinson's disease (PD), and in healthy volunteers (n = 19, 38, and 13, respectively). A mean template was created from coregistered control studies. Registration errors were assessed using studies with simulated binding deficits. Patient studies were registered to the mean template, and striatal binding was calculated from a corresponding map of 3-dimensional regions of interest (ROIs). The striatal binding ratio and deficits determined by voxelwise comparison with the normal template were investigated and tested with various 3-dimensional ROI sizes and positions. Separation of patient groups was determined by t score after automatically processing all studies. Results were compared with manual ROI analyses. **Results:** The automatic method was completely reproducible in 64 of 70 cases. The best diagnostic discriminator was the minimum binding ratio of the 2 striatal nuclei, with the following values: NIPS, 1.33 ± 0.13; PD, 1.50 ± 0.12; healthy volunteers, 1.49 ± 0.08 (±SD). The deficit size from voxelwise analysis was: NIPS, 20.5 ± 8.2 mL; PD, 9.5 ± 8.3; healthy volunteers, 8.9 ± 6.0 (±SD). The accuracy, measured by receiver operating characteristic areas, was 0.85 ± 0.05, 0.77 ± 0.06, and 0.80 ± 0.06 (±SE) for the optimal predictor (automated) and 2 blinded observers (manual), respectively. **Conclusion:** A new 3-dimensional, automated technique has been developed to semiquantitate receptor density that dramatically improves reproducibility. The optimal diagnostic discriminator of parkinsonism determined by the automatic technique has good accuracy compared with the manual technique.

Key Words: automatic; registration; dopamine imaging; Parkinson's disease; reproducibility; receptors; SPECT

J Nucl Med 2000; 41:220–227

Imaging of ^{123}I-iodobenzamide (IBZM) uptake has been shown to be useful for the semiquantitative analysis of dopamine (D_2) receptor density (1–3) and has been applied to the differential diagnosis of parkinsonism. Parkinson's disease (PD) is the diagnosis in most patients presenting with parkinsonism, but approximately 10% of parkinsonism arises from other neurodegenerative diseases such as multiple system atrophy (4) and progressive supranuclear palsy (5). The latter nonidiopathic parkinsonism (NIPS) disorders have been distinguished from PD by ^{123}I-IBZM imaging with SPECT, which has demonstrated a lower specific uptake in the striatum for the NIPS group than in either PD or control groups (6–9).

The variety of manual, operator-intensive techniques of analysis that have been applied semiquantitatively to ^{123}I-IBZM images have shortcomings. Both inter- and intraobserver variability can arise in manual semiquantitation (10) as a result of reliance on an operator to realign the studies to a standard orientation and draw regions of interest (ROIs), and this variability hampers the development of an optimal analysis method. Also, most manual techniques select a single slice of the tomographic study for analysis, and therefore do not use the full information of a 3-dimensional SPECT study. This study presents an automatic technique of quantifying ^{123}I-IBZM binding and compares it with a standard manual technique, as applied to the discrimination of parkinsonism.

MATERIALS AND METHODS

Subjects

Thirty-eight patients (16 women, 22 men) diagnosed clinically with PD and 19 patients (7 women, 12 men) with NIPS were included in this retrospective study. These patients had not been treated with dopaminomimetic drugs during the 2 d before the receptor imaging. The mean age of the PD patients was 57 ± 13 y (range, 28–82 y), whereas for the NIPS patients it was 60 ± 14 y (range, 29–78 y). The control group comprised 13 volunteers (4 women, 9 men) who were free of neurological diseases. This group had a mean age of 54 ± 11 y (range, 35–68 y). There were no significant differences in the ages of these 3 groups (t test and F test,

$P > 0.10$). The clinical diagnosis of PD and NIPS patients was based on a positive (PD) or negative (NIPS) response to the apomorphine test (*11*) and/or dopamine replacement therapy (*12*). All patients gave informed consent before undergoing ^{123}I-IBZM SPECT imaging for routine clinical diagnosis.

SPECT Imaging

The SPECT studies were acquired with a triple-head Siemens Multispect 3 (Chicago, IL) camera using special high-resolution ^{123}I collimators (*13*). Acquisitions began 2 h ± 5 min after injection of 185 MBq ^{123}I-(*S*)-2-hydroxy-3-iodo-6-methoxy-*N*-[1-ethyl-2-pyrrolidinyl-methyl]-benzamide (IBZM). This radiopharmaceutical is commercially available from Amersham-Cygne (Eindhoven, The Netherlands) and has specific activity 185–370 TBq/mmol. The matrix was 128 × 128, with a voxel size of 3.56 mm. The camera heads followed a circular orbit of about 13-cm radius over 120° and 30 views, scanning for 80 s/view, resulting in 90 projections, for a total scan time of 42 min. The counting rate in a projection was 2.5–3.0 × 10⁴ counts, depending on the patient. Technologists visually inspected the sinograms for motion and manually corrected individual projections as required; studies that could not be corrected were discarded. The reconstruction was performed with filtered backprojection (Butterworth filter of seventh order, cutoff 0.38 Nyquist corresponding to 0.53 cycles/cm), followed by Chang attenuation correction using a coefficient of 0.1/cm.

Software

Image registration and quantification were performed with the program BRASS (Nuclear Diagnostics, Stockholm, Sweden), which had been modified by the author for this purpose. Statistics were calculated using Statistica software (Tulsa, OK). Receiver operating characteristic (ROC) curves were generated with Medcalc (*14*) software (Medcalc Software, Mariakerke, Belgium).

Registration

Control studies were coregistered to develop normal templates, i.e., maps of the mean and SD values of all voxels. Subsequently, individual patient cases were registered to the mean template and compared on a voxel-by-voxel basis.

The registration method adjusted 9 parameters, 3 each for rotation, translation, and anisotropic scaling. For registration and all subsequent analysis, a 64 × 64 image matrix that contained the brain volume was extracted from the 128 × 128 reconstruction matrix, without changing the original voxel size.

The coregistration of ^{123}I-IBZM studies occurred in steps. The first was fast, geometric alignment with automatic detection of the basal ganglia as markers. Anisotropic scaling parameters were found using the principal axes technique (*15*). The midpoint between the striatal nuclei of study and template was matched to calculate the translation. The rotation angles were found by the principal axes method, although only the anterior basal ganglia was used for this determination, because the anterior is less often asymmetrically affected by NIPS than the posterior.

Second, masking was applied to the approximately fitted study to eliminate the influence of facial activity and external noise. A mask was formed from template voxels below a 35% threshold. The approximate fit of the study to the template in the previous step ensured that these voxels constituted a valid mask boundary.

Third, an iterative, simplex algorithm (*16*) was applied twice to search for the set of 9 transform parameters that gave the minimum count difference between the masked study and the template. During the first application, only voxels with a count value above the 35% threshold were allowed to contribute to the sum of absolute count difference (SAD), and nearest-neighbor interpolation (*16*) was used to transform the study at each iteration step. The second application did not affect the scaling, used SAD with a 65% threshold (in effect, ignoring the background), and calculated transformations with trilinear interpolation. In each case, the search for the transform stopped when the SAD changed by <0.1%. After the determination of the transforms in each step, these were combined into a single transform product and applied to the original reconstructed study to minimize interpolation error.

The automated alignment of NIPS or PD studies was carefully assessed visually (Fig. 1), by examining the difference between the normalized, registered patient image and the template. If automated registration was not deemed adequate, the study was manually aligned to the normal template, and the resultant image was used for subsequent automatic quantification.
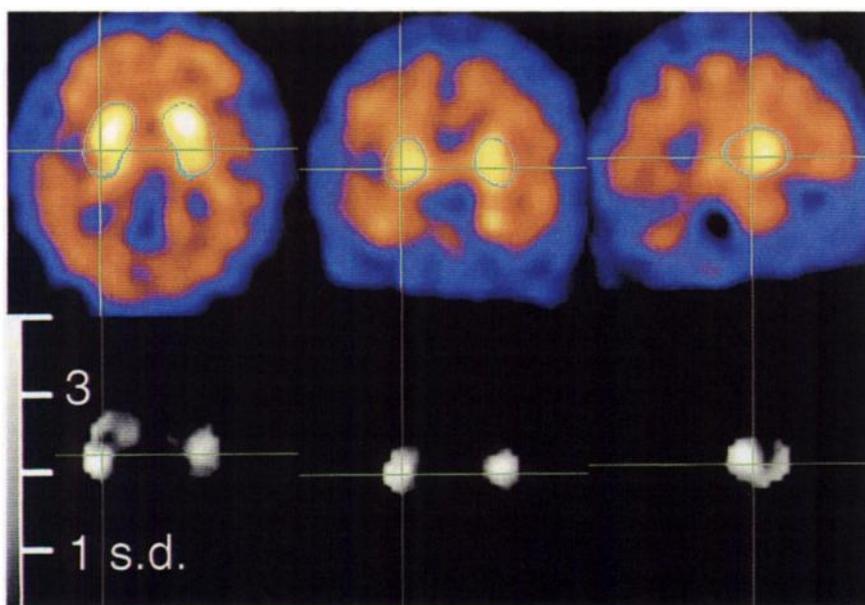
FIGURE 1. NIPS patient's SPECT study, with transverse, coronal, and sagittal views (left to right, top row). Bottom row is corresponding Z-score map with pixel brightness indicating abnormality. Area of basal ganglia defined by template is outlined by blue isocontour, in which threshold was adjustable. Green cross is display device used to locate voxel, synchronized in all views. These devices aided visual assessment of registration accuracy.

Normal Template

The 19 studies from healthy volunteers were registered to form a normal template. This template provided mean and SD values for each voxel and for specific 3-dimensional volumes defined by a fixed volume-of-interest (VOI) map. The method of template creation is similar to that previously described (17,18), except with the addition of the above-mentioned registration technique. The template was created as follows:

1. A representative study from a healthy volunteer was selected, with little facial activity and very normal-appearing striatal nuclei. This study was realigned so that the orbitomeatal line was parallel to the transverse slices.
2. All other healthy volunteer studies were registered to the reference study selected in step 1, and the registration was visually assessed for accuracy.
3. Control studies were normalized by their maximum count values to the reference study, and the set of means calculated voxelwise formed a normal template volume.
4. The control studies were normalized by their maximum count values to the maximum of the template determined in step 3.
5. The mean and coefficient of variation were calculated voxelwise, with results placed in the corresponding voxels of the mean and variation templates, respectively.
6. Steps 1 to 5 were repeated with the mean template taking the role of the reference study, to reduce any bias resulting from the particular reference study chosen in step 1.

The mean template that resulted from this process is shown in Figure 2. Individual patient studies are less smooth than the mean, which is reflected in the variation template. These templates were used to calculate a Z-score map for each patient study registered to the mean template.

VOI Map

The anatomical volumes chosen for quantification are defined by the VOI map (Fig. 2), developed by choosing specific thresholds of the template as boundaries. A fixed VOI map was feasible for quantification because the registration process matches the study to both template and corresponding VOI map. The basal ganglia VOIs included areas of the template above 65% of the maximum. Each striatal nucleus was also subdivided into caudate and putamen by a plane perpendicularly transecting the transverse slices, and positioned one third of the distance from the anterior to posterior tips of the striatal nucleus. All VOIs were the depth of the basal ganglia (7 slices × 3.56 mm). The map was made bilaterally symmetrical. The quantification regions could be interactively altered within the software (e.g., to reduce the number of slices, alter the threshold boundaries, or combine regions), so that various parameters could be tested to optimize the automatic quantification.

Quantification

Manual. For comparison with the automatic technique, each case study was analyzed with the semiquantitative manual technique that has been used in the routine, clinical interpretation of ^{123}I-IBZM images (19). The transverse slices were aligned parallel to the orbitomeatal line. On the ^{123}I-IBZM transverse slice containing the image maximum was drawn an ROI of a fixed size of 50 ± 3 voxels for each striatal nucleus, and an irregular frontal cortex region (~40% isocontour). From these ROIs the ratio of the mean counts in the basal ganglia to that of the frontal cortex (BG/FC) was calculated. This technique was applied by 3 observers. Observers 1 and 2 quantified all studies (n = 70), whereas observer 3 quantified 54 (excluding the recently acquired 5 NIPS and 11 healthy volunteers). Observers 2 and 3 were blinded, whereas observer 1 performed the quantification during the clinical routine and thus was unblinded.

Automatic. The patient images registered to the template were quantitated by a fully automated algorithm. First, the following binding ratios were tested: (a) BG/FC, (b) basal ganglia-to-occipital cortex binding (BG/OC), (c) posterior (putamen) basal ganglia-to-frontal cortex, (d) anterior (caudate) basal ganglia-to-frontal cortex, and (e) lower of the 2 nuclear BG/FC ratios (minBG/FC).

Second, the striatal absolute asymmetry was quantified, and defined as:

$$\text{Asymmetry} = 100\% \times |L - R|/(0.5 \times (L + R)),$$

where L and R are the mean count values in the left and right striatal VOIs, respectively. The absolute asymmetry index is not a

FIGURE 2. Mean template for control studies from healthy volunteers is shown in top row, and matched VOI map in bottom row. Transverse, coronal, and sagittal views are shown from left to right. Mean template was composed of 13 registered control studies. Map VOIs are posterior basal ganglia (A, B), anterior basal ganglia (C, D), frontal background (E), and occipital background (F). Blue curve is 35% isocontour of template.
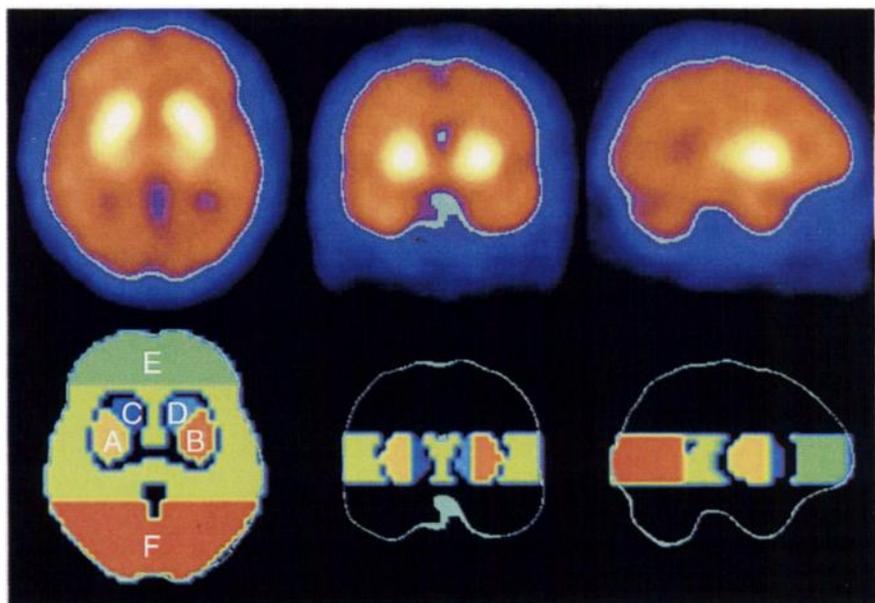
normal variate; therefore, differences therein were tested with the nonparametric Mann–Whitney U test.

Third, the abnormal region size and severity were calculated by the voxel-based region-growing technique. This technique "grows" regions of contiguous voxels that have abnormally low Z-scores within the striatum. The severity is defined as the average of the voxel Z-scores in the grown regions.

The diagnostic efficacy of a parameter (e.g., the number of basal ganglia slices included for the calculation of BG/FC) was estimated from the degree of separation of the PD and NIPS groups. The group difference was almost always significant. Therefore the relevant question was the parameter-level that gave the greatest separation of the PD and NIPS groups. By the automatic method it was possible to quantify the entire set of studies, calculate the PD–NIPS group separation, and then repeat this step for a series of parameter levels (e.g., VOI size). The parameters associated with the largest t value for the PD–NIPS difference were accepted as optimal for this population.

The optimization procedure was applied to several parameters, for example, the number of slices comprised by the basal ganglia VOIs (varied from 1 to 7). The boundaries of the VOIs were modified by thresholds to discard voxels with low counts (i.e., low signal-to-noise ratio). The thresholds tested were: a percentage of the image maximum, a percentage of the frontal cortex maximum, and a strict count threshold.

Accuracy was assessed by ROC curve analysis (20–23) of the values from the optimal discriminator minBG/FC derived from automatic quantification, and manually quantified BG/FC. Healthy volunteer and PD groups were combined to increase the power of the ROC analysis. The entire set of studies (n = 70) was used to generate these curves, with the exception of some studies not performed by observer 3 (n = 54, excluding the recently acquired 5 NIPS and 11 controls).

Reproducibility of Automatic Technique

To estimate the error in the automated fitting for studies with binding deficits, 6 simulated defects were generated on 5 normal studies. Each artificial defect either covered the entire striatal nucleus (bilateral or unilateral) or bilaterally covered the putamen. Each defect was created by reducing the striatal counts voxelwise by a fraction (25% or 55%) of the previous striatum-specific binding. Images with simulated defects and the corresponding controls were then misaligned and subsequently registered to the normal template. Twelve misalignments were tested for each study for a total of 7 × 5 × 12 = 420 registrations. Each misalignment affected 6 of the 9 parameters (±5 pixel shift, ±10% scaling, or ±20°). The registration parameters from the control study without defects provided an absolute reference to which the registrations of images with simulated defects were compared. The registration error for each transformation parameter was defined as the change in that parameter (averaged across misalignment trials) related to the introduction of the defect, calculated individually for each study. The reproducibility of the registration was defined as the SD of each transformation parameter across misalignment trials for the control studies.

Similarly, the error in quantification was estimated by means of the misalignments and defects described earlier. The error was defined as the change in BG/FC when the study had been fitted with or without the defect present. The defect was present in both cases for the quantification, but not for the fitting. The reproducibility of

BG/FC was defined as the SD of BG/FC across misalignment trials among control studies (without defects).

RESULTS

Reproducibility of Manual Technique

The interobserver variation of BG/FC is shown in Figure 3, a Bland–Altman plot (24) of the SD versus the mean for each study, as calculated by the 3 observers. The average across all studies of the BG/FC variation was 0.04 (SD). Using a BG/FC threshold of 1.6, the interobserver variation caused 8–18 of 70 classification disagreements between observer pairs.

Reproducibility of Automatic Technique

The reproducibility of the registration parameters, tested with a series of 12 misalignments for 5 control studies, was 0.2 pixel shift, 1.1% scaling, and 2.6° angle (SD). The reproducibility of BG/FC was 0.01 (SD).

The registration errors that arose from the introduction of artificial defects are listed in Table 1. The registration errors for the moderate (55%) defects were smaller than those for the severe (25%) defects, and hence were omitted from Table 1. The table illustrates that there was no systematic error as a result of the unilateral defect; a 1-pixel shift as a result of the bilateral putamen defect; and a 1-pixel shift and 10% minification scaling error that resulted from the bilateral striatal defect. The quantification error of BG/FC resulting from fitting with the severe (25%) defect was 0.00 ± 0.01 (unilateral striatum), 0.03 ± 0.02 (bilateral putamen), and −0.01 ± 0.01 (bilateral striatum).

Automatic Registration

The success rate of automatic fitting was 64 of 70 (92%). All control studies were fitted correctly. The 6 studies that were not fitted correctly included 4 of 38 that were PD and 2
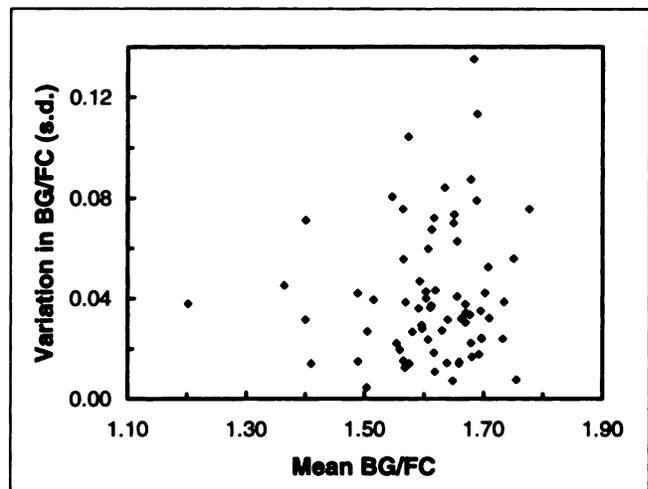
FIGURE 3. Manual quantification interobserver variability for entire population of PD, NIPS, and healthy volunteers. SD of binding ratio (BG/FC) for 3 observers is plotted against mean binding ratio for each patient study. Binding ratio was measured by observers 1 (n = 70), 2 (n = 70), and 3 (n = 54).

TABLE 1
Error in Registration After Misalignments, Demonstrating Effect of 3 Artificial Receptor-Density Defects*

	Unilateral striatum	Bilateral putamen	Bilateral striatum
X shift (mm)	0.0 ± 0.3	0.1 ± 0.1	0.2 ± 0.1
Y shift (mm)	0.0 ± 0.2	−1.0 ± 0.3†	0.6 ± 0.2†
Z shift (mm)	0.2 ± 0.2	0.0 ± 0.2	0.0 ± 0.0
X scale (%)	−0.6 ± 0.6	−0.1 ± 0.2	−8.1 ± 1.0†
Y scale (%)	−0.3 ± 0.7	−0.2 ± 0.2	−7.9 ± 0.9†
Z scale (%)	−1.0 ± 1.2	−0.1 ± 0.5	−14.2 ± 1.7†
Yaw angle (°)	1.3 ± 0.8	0.8 ± 0.9	2.2 ± 2.0
Roll angle (°)	0.7 ± 1.3	0.1 ± 0.9	2.3 ± 1.7
Pitch angle (°)	−0.8 ± 0.4	0.6 ± 0.6	−1.6 ± 1.3

*All defects were 25% of original striatal-specific receptor density. Mean error ± SD. n = 5 subjects, each misaligned 12 times. Errors were calculated individually for each subject. The X, Y, and Z labels refer to lateral, anterior-posterior, and superior-inferior directions, respectively. Yaw, roll, and pitch are rotations elicited by head shaking "no," tilting the side of the head toward the shoulder, and nodding "yes," respectively.

†Errors significantly different from 0 (P < 0.05, ANOVA with post hoc Newman–Keul test).

of 19 that were NIPS. In all cases the error seemed small, as supported by the determination that the BG/FC ratio changed, at most, by 0.01 when manual fitting was substituted for automatic fitting.

Optimizing the Diagnostic Discriminator by the Automatic Technique

Number of VOI Slices. Increasing the number of slices in each VOI reduced the BG/FC ratio for all studies. Using a VOI encompassing the basal ganglia (7 slices) did improve the group separation slightly (1 slice, $t = 4.7$; 7 slices, $t = 4.8$). All tests of the differences in the BG/FC means of the NIPS and PD groups were highly significant ($P < 0.0001$). The use of VOIs, including the entire basal ganglia rather than a single slice, was supported because, in cases of disagreement, inspection revealed that the maxima of the striatal nuclei were not contained in a single slice, leading to incorrect results. All subsequent results had 7 slices in each VOI.

Background Region Threshold. Three threshold types were examined, each optimized by testing a series of at least 4 thresholds. First, a fixed-count threshold (60) applied to the frontal region gave the greatest separation of the PD and NIPS groups ($t = 4.8$). Second, a threshold as percentage of the study maximum was tested, similar to the contour that an observer might use when drawing an ROI, and by this method a 45% threshold ($t = 3.7$) was best. The third threshold type was computed as a percentage of the maximum within the frontal reference region, for which 85% ($t = 3.4$) was best. The optimal values for each threshold method were then selected for comparison by analysis of variance. In Figures 4A–C, the mean BG/FC values of the NIPS, PD,
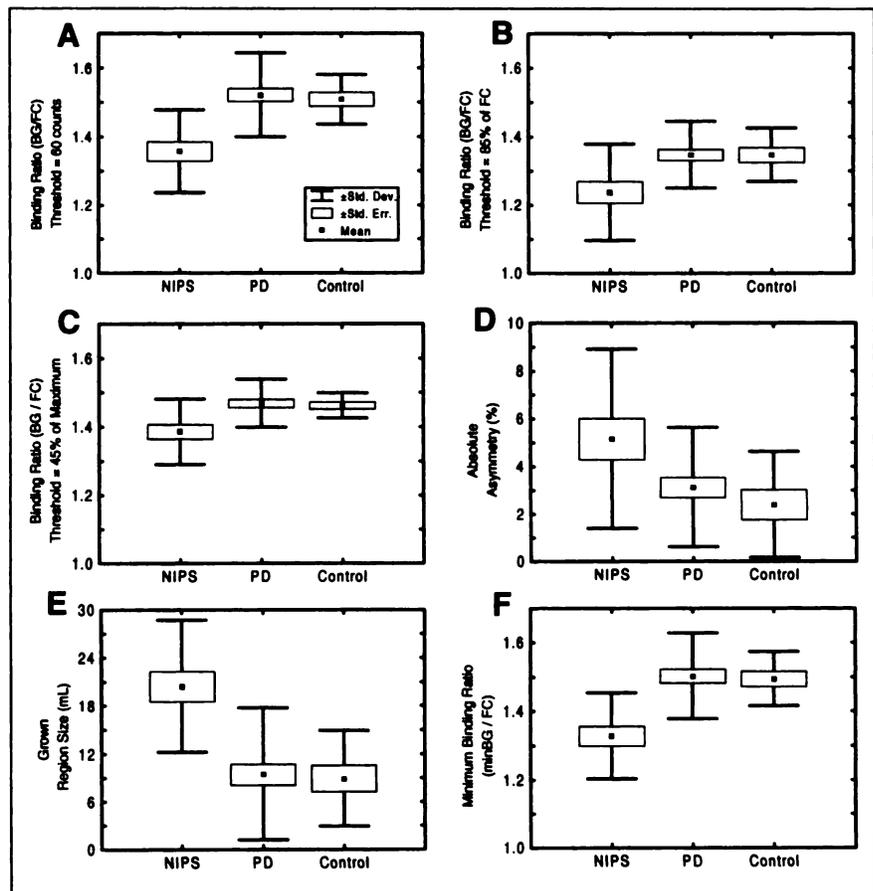
FIGURE 4. Comparison of discrimination methods derived from automatic technique. Each plot shows NIPS, PD, and healthy volunteer means, from left to right, calculated from all cases. (A) BG/FC, with constant threshold = 60 counts; (B) BG/FC, with threshold = 85% of FC region maximum. (C) BG/FC, with threshold = 45% of study maximum. (D) Absolute asymmetry (%) of basal ganglia nuclei. (E) Abnormal region size (mL), determined by voxelwise comparison with template. Size is sum of regions found in both striatal nuclei. (F) Minimum BG/FC of striatal nuclei.

and healthy volunteer groups are plotted for each of the 3 threshold methods. Using the fixed-count threshold applied to the reference region provided the largest separation of the NIPS group from PD ($t = 4.8$) and healthy volunteers ($t = 4.0$), and was subsequently adopted to calculate the binding ratio.

Asymmetry. There was a marginal significant difference between the asymmetry of the NIPS group and that of either the PD or healthy volunteer groups ($P = 0.04$ in both cases), whereas the PD and healthy volunteer groups did not differ ($P > 0.05$). These results are shown in Figure 4D.

Region Growing. The total size and severity (i.e., average Z-score of the voxels) of grown regions were calculated. With the product of severity and size as the index of abnormality, an SD threshold of 0.5 produced the best separation of the NIPS and PD groups ($t = 4.4$). Adopting the 0.5 SD threshold, the severity values were 2.7 ± 0.2, 1.8 ± 0.1, and 1.6 ± 0.1 for the NIPS, PD, and healthy volunteer groups, respectively. The summed sizes of the regions are shown in Figure 4E. The region size was the superior discriminator of the NIPS and PD groups (size, $t = 4.7$; severity, $t = 3.9$).

Binding Ratios. All pairs of binding ratios had high correlations, and in particular BG/FC was correlated to BG/OC by $r = 0.95$. The BG/OC values were systematically 4% higher than BG/FC ($P < 0.0001$). The minBG/FC parameter had the highest t values in either the test of NIPS against PD or healthy volunteer groups, although only marginally better than those obtained with BG/FC. The means of the minBG/FC parameter are shown in Figure 4F. By analysis of variance, it was determined that the difference between the mean minBG/FC of the NIPS group and either the PD or healthy volunteer group was significant ($P < 1 \times 10^{-5}$). The PD group had a slightly larger average minBG/FC than the healthy volunteers, but this was not significant ($P = 0.05$).

ROC Analysis. ROC curves were determined for the manual technique using BG/FC values and for the automatic technique using the optimal discriminator, minBG/FC. The ROC curves have underlying areas (±SE) of 0.85 ± 0.05, 0.88 ± 0.04, 0.77 ± 0.06, and 0.80 ± 0.06, respectively, for the automatic technique, and observers 1, 2, and 3 using the manual technique (Fig. 5). These results suggest that the automatic technique's effectiveness lies between that of the unblinded observer 1 and the blinded observers (2 and 3). However, with this low number of studies, the differences in the areas cannot be distinguished statistically ($P > 0.05$).

DISCUSSION

Because in recent years dopamine D_2 receptor imaging with SPECT has been shown to provide valuable diagnostic information in a variety of neuropsychiatric disorders, there is increasing interest in developing observer-independent techniques of image analysis like those developed for PET studies (25). The rationale is not only an improvement in the primary diagnosis of various disease entities, but also the
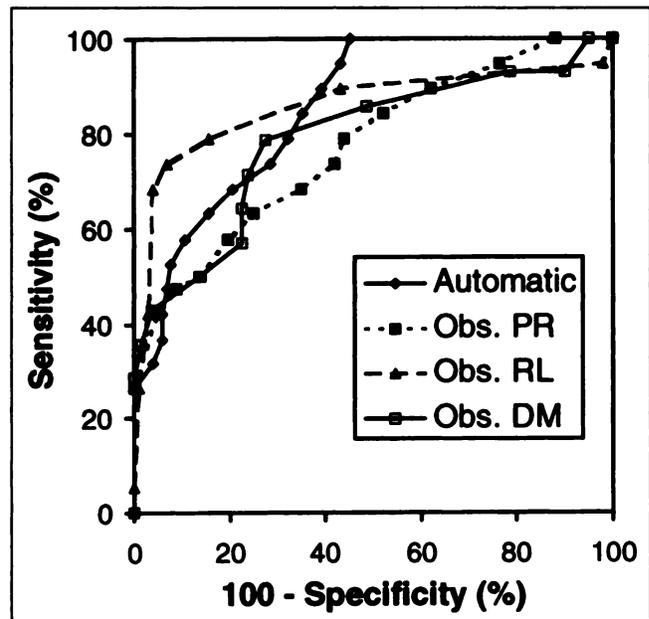
FIGURE 5. ROC curves for automatic technique and manual technique (unblinded observer 1 and blinded observers 2 and 3). Positive cases: NIPS; negative cases: PD and healthy volunteers. All studies (n = 70) were used to generate these curves except for curve of observer 3 (n = 54, missing 5 NIPS and 11 healthy volunteers).

provision of an objective means for intraindividual follow-up studies to detect subtle changes resulting from disease progression or the beneficial effects of putative neuroprotective drugs. In this study, we present an automatic technique of quantifying IBZM binding and compare it with an established manual technique, as applied to the differential diagnosis of parkinsonism.

The automatic technique was completely reproducible in 64 of 70 cases in which automatic registration was successful. When combined with automatic quantitation, the inter- and intraobserver variability is completely eliminated. Incorrect fits seemed to occur because of (a) low counts throughout the study, (b) prominent facial activity, and (c) abnormal striatal nucleus shape that could not be normalized by anisotropic scaling. The assessment of reproducibility using artificial defects demonstrated that the automated fitting was very robust and the errors had a small effect on the binding ratio. The severe, bilateral artificial defect tested was equivalent to the NIPS study with the lowest binding ratio in our population, and thus the errors are an upper estimate. An advantage of the automated fitting technique is that it does not require coregistered MRI studies or external markers, although in future work it would be worthwhile to further validate the fitting by these methods.

This study provides an estimate of the interobserver variability for a particular technique of manual quantification for comparison with the automated technique. The interobserver variability of the manual technique was significant, causing 8–18 of 70 disagreements in classification. The most likely causes of interobserver discrepancies were the

choice of the background ROI, especially when it had low counts, and the choice of the transverse plane, which had a dramatic effect when the uptake was concentrated in a small part of the striatum. Overcoming interobserver variability, as well as the harmonization of acquisition protocols, is required to determine the optimal analytical method for a multicenter study. These results are not a definitive assessment of manual quantification variability, which would consider intraobserver variability, training effects, and other possible manual techniques, and hence is beyond the scope of this investigation.

The analysis of the automated quantification results demonstrated that the choice of background threshold had a large effect on the separation of the PD and NIPS groups compared with other variables. The fixed-count threshold that proved to be best reduces the influence of noisy, low-count voxels, suggesting that other more sophisticated noise-suppression techniques may further improve the diagnostic value of ^{123}I-IBZM studies.

Of the variety of diagnostic discriminators tested, the binding ratio was best. Using the minimum binding ratio of the 2 nuclei (minBG/FC) was marginally better than other binding ratios, probably because it capitalized on the greater asymmetry of the NIPS studies.

The procedure of optimizing the thresholds and other parameters affecting the quantification results was made possible by the processing efficiency and reproducibility of the automated technique. Although the resultant parameters are dependent on the acquisition setup, filters, camera, and other factors, this does not diminish their value for future work using the same protocol. Moreover, these parameters may provide a starting point for those using the manual technique and lacking the resources to conduct an optimization study for their own protocols.

The voxel-based region growing technique did not provide better diagnostic information than the binding ratio. The separation of the NIPS and PD groups, using the size of the grown regions, was nevertheless similar to the separation achieved by using the binding ratio. With a larger group of healthy volunteers (e.g., 20) to form the template, the estimation of voxel statistics would improve and could make the region-growing technique superior to binding ratios. Voxelwise evaluation has the potential to assess the substriatal uptake and thereby elucidate the underlying pathophysiology more clearly than regional analysis. The value of voxelwise analysis would be increased with coregistered MRI studies to accurately delineate the anatomy corresponding to receptor deficits.

The ROC curve analysis qualitatively suggests that the automatic technique is more effective than the blinded observers 2 or 3, but less effective than unblinded observer 1. The predictive power of the automatic technique could not be distinguished statistically from any observer using the manual technique because of the small population sample. In our future work, we would like to validate the automated technique with a larger set of studies in each group. The

discrimination may also be improved by using a more uniform NIPS group (e.g., only multiple system atrophy cases). An important limitation to our assessment of the diagnostic accuracy of the IBZM studies is that the true diagnoses are not known with certainty, because approximately 20% of patients clinically diagnosed as having PD have other neuropathologic diseases (26). Therefore, it would be valuable to optimize the automated technique with studies that also have postmortem confirmation of the diagnoses. Nevertheless, the uncertainty in clinical diagnoses should equally affect the manual and automatic methods without biasing that comparison.

The technique presented in this study has relevance beyond the present application to the diagnosis of parkinsonism by ^{123}I-IBZM studies. By a similar technique, other receptor studies might be automatically aligned and quantified, to practically eliminate inter- and intraobserver variability. In the future, we would like to combine the quantification results from ^{123}I-IBZM studies with those from presynaptic dopamine transporter studies to synergistically improve the understanding of the pathophysiology underlying parkinsonism. The reproducibility of this technique is important for both optimizing the differential diagnosis and detecting changes among intraindividual follow-up studies.

CONCLUSION

This study demonstrates a new automatic technique of aligning and quantifying ^{123}I-IBZM studies that greatly improves reproducibility without degrading diagnostic accuracy. The promising nature of these preliminary results requires validation with a large trial study.

REFERENCES

1. Kung HF, Pan S, Kung MP, et al. In vitro and in vivo evaluation of ^{123}I-IBZM: a potential CNS D-2 dopamine receptor imaging agent. J Nucl Med. 1989;30: 88–92.
2. Costa DC, Verhoeff NPLG, Cullum ID, et al. In vivo characterisation of 3-iodo-6-methoxybenzamide 123-I in humans. Eur J Nucl Med. 1990;16:813–816.
3. Bruecke T, Wenger S, Asenbaum S, et al. Dopamine D2 receptor imaging and measurement with SPECT. Adv Neurol. 1993;60:494–500.
4. Quinn N. Multiple system atrophy: the nature of the beast. J Neurol Neurosurg Psychiatry. 1989;52(suppl):78–89.
5. Litvan I, Campbell G, Mangone CA, et al. Which clinical features differentiate progressive supranuclear palsy (Steele-Richardson-Olszewski syndrome) from related disorders? Brain. 1997;120:65–74.
6. Schulz JB, Klockgether T, Petersen D, et al. Multiple system atrophy: natural history, MRI morphology and dopamine receptor imaging with [123I]IBZM SPECT. J Neurol Neurosurg Psychiatry. 1994;57:1047–1056.
7. van Royen E, Verhoeff NF, Speelman JD, Wolters EC, Kuiper MA, Janssen AG. Multiple system atrophy and progressive supranuclear palsy: diminished striatal

dopamine-D2 receptor activity demonstrated by [¹²³I]IBZM single-photon emission computed tomography. *Arch Neurol.* 1993;50:513–516.

8. Cordes M, Hierholzer J, Schelosky L, et al. IBZM-SPECT imaging in Parkinson's disease. *Adv Neurol.* 1993;60:525–528.

9. Hertel A, Weppner M, Baas H, et al. Quantification of IBZM dopamine receptor SPET in de novo Parkinson patients before and during therapy. *Nucl Med Commun.* 1997;18:811–822.

10. Verhoeff NP, Kapucu O, Sokole-Busemann E, van Royen EA, Janssen AG. Estimation of dopamine D2 receptor binding potential in the striatum with iodine-123-IBZM SPECT: technical and interobserver variability. *J Nucl Med.* 1993;34:2076–2084.

11. Hughes AJ, Lees AJ, Stern GM. Challenge test to predict the dopaminergic response. *Neurology.* 1991;41:1723–1725.

12. Kempster PA, Frankel JP, Stern GM, Lees AJ. Comparison of motor response to apomorphine and levodopa in Parkinson's disease. *J Neurol Neurosurg Psychiatry.* 1990;53:1004–1007.

13. Muenzing W, Raithel E, Tatsch K, Hahn K. Design of a special collimator for iodine-123 and first experiences in clinical practice [abstract]. *Nuklearmedizin.* 1996;35:A48.

14. Schoonjans F, Zalata A, Depuydt CE, Comhaire FH. MedCalc: a new computer program for medical statistics. *Comput Methods Programs Biomed.* 1995;48:257–262.

15. Alpert NM, Bradshaw JF, Kennedy D, Correia JA. The principal axes transformation—a method for image registration. *J Nucl Med.* 1990;31:626–633.

16. Press WH, Teukolsky SA, Vetterling WT, Flannery BP. *Numerical Recipes in C.* 2nd ed. New York, NY: Cambridge University Press; 1992.

17. Slomka PJ, Hurwitz G, Stephenson JA, Cradduck TD. Automated alignment and sizing of myocardial stress and rest scans to three-dimensional normal templates using an image registration algorithm: a method for reproducible quantification. *J Nucl Med.* 1995;36:1115–1122.

18. Slomka PJ, Stephenson J, Reid R, Hurwitz GA. Automated template-based quantification of brain SPECT. In: De Deyn PP, ed. *Textbook of SPECT in Neurology and Psychiatry.* London, UK: BIOS Scientific; 1998:507–512.

19. Schwarz J, Oertel WH, Tatsch K. Iodine-123-iodobenzamide binding in Parkinsonism: reduction by dopamine agonists but not L-dopa. *J Nucl Med.* 1996;37:1112–1115.

20. Lusted LB. Decision-making studies in patient management. *N Engl J Med.* 1971;284:416–424.

21. Goodenough DJ, Rossmann K, Lusted LB. Radiographic applications of receiver operating characteristic (ROC) curves. *Radiology.* 1974;110:89–95.

22. Swets JA. ROC analysis applied to the evaluation of medical imaging techniques. *Invest Radiol.* 1979;14:109–121.

23. Hanley J, McNeil B. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology.* 1982;143:29–36.

24. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet.* 1986;1:307–310.

25. Minoshima S, Frey KA, Koeppe RA, Foster NL, Kuhl DE. A diagnostic approach in Alzheimer's disease using three-dimensional stereotactic surface projections of fluorine-18-FDG PET. *J Nucl Med.* 1995;36:1238–1248.

26. Jellinger K. The pathology of parkinsonism. In: Marsden DC, Fahn S, eds. *Movement Disorders.* Vol. 2. London, UK: Butterworths; 1987:124–165.