Movement Disorders
Vol. 18, No. 1, 2003, pp. 70–80

Automatic Assessment of Levodopa-Induced Dyskinesias in Daily Life by Neural Networks

Noël L.W. Keijsers, MSc,[1]* Martin W.I.M. Horstink, MD, PhD,[2] and Stan C.A.M. Gielen, PhD[1]

[1]*Department of Biophysics, University of Nijmegen, Nijmegen, The Netherlands*
[2]*Department of Neurology, University of Nijmegen, Nijmegen, The Netherlands*

Abstract: We developed an objective and automatic procedure to assess the severity of levodopa-induced dyskinesia (LID) in patients with Parkinson's disease during daily life activities. Thirteen patients were continuously monitored in a home-like situation for a period of approximately 2.5 hours. During this time period, the patients performed approximately 35 functional daily life activities. Behavior of the patients was measured using triaxial accelerometers, which were placed at six different positions on the body. A neural network was trained to assess the severity of LID using various variables of the accelerometer signals. Neural network scores were compared with the assessment by physicians, who evaluated the continuously videotaped behavior of the patients off-line. The neural network correctly classified dyskinesia or the absence of dyskinesia in 15-minute intervals in 93.7, 99.7, and 97.0% for the arm, trunk, and leg, respectively. In the few cases of misclassification, the rating by the neural network was in the class next to that indicated by the physicians using the AIMS score (scale 0–4). Analysis of the neural networks revealed several new variables, which are relevant for assessing the severity of LID. The results indicate that the neural network can accurately assess the severity of LID and could distinguish LID from voluntary movements in daily life situations. © 2002 Movement Disorder Society

Key words: automatic assessment; activities of daily living; Parkinson's disease; levodopa-induced dyskinesia; accelerometers; neural networks

Levodopa-induced dyskinesia (LID) is a disabling and distressing complication of chronic levodopa therapy in patients with Parkinson's disease.[1,2] Therefore, new pharmacological and surgical treatments to reduce these dyskinesias are of increasing interest.[3–5] To evaluate medication and surgical treatment, it is important that dyskinesia can be assessed objectively in daily life. However, the commonly used methods to assess LID have several limitations.[6–8] For example, long-term assessment by experts is not feasible as a routine procedure, and self-assessment of LID by patients can be unreliable.[9,10] Moreover, the ratings are subjective. For these reasons, a portable device that can assess LID automatically and objectively in daily life would be highly useful.[11]

*Correspondence to: Noël L.W. Keijsers, MSc, Department of Biophysics UMC, BEG 231, University of Nijmegen, 6525 EZ Nijmegen, The Netherlands. E-mail: noelk@mbfys.kun.nl

Received 25 March 2002; Revised 29 June 2002; Accepted 9 July 2002

Recently, several studies attempted to establish an objective and automatic method to assess dyskinesia using accelerometers, which can measure movements of patients without any discomfort.[12–15] Burkhard and colleagues[12] used a rotation-sensitive movement monitor (RoMM) and could successfully quantify and characterize dyskinesia for patients who were asked to abstain from voluntary movements. In a study by Hoff and associates,[13] patients were tested in a set of seven tasks of 1-minute duration each. These authors used a linear discriminant analysis and could assess the severity of LID for tasks in which patients abstained from voluntary movements. However, they had problems in assessing LID when voluntary movements were present, such as during drinking and walking. Keijsers and coworkers[14] used the same data set as that used in the study by Hoff and colleagues[13] but used neural networks instead of linear discriminant analysis to assess LID. The neural-network approach showed a better performance than the linear classification technique used by Hoff and associates,[13] and also appeared to better distinguish between

LID and voluntary movements. However, the results of Keijsers and coworkers[14] were far from optimal, indicating that considerable improvement is needed to obtain a reliable method which can be used in daily life. In another study, Manson and colleagues[15] attached a triaxial accelerometer to the shoulder and showed that the accelerations in the 1 to 3 Hz frequency band correlated well with the AIMS scale[16] for various tasks. Like Keijsers and associates,[14] Manson and coworkers[15] were able to assess the severity of LID, even when patients made voluntary movements. However, a main limitation in the study by Manson and colleagues[15] may be the low specificity for mild dyskinesias. Because all patients in that study suffered from severe dyskinesia during the test, the method was not validated to assess mild dyskinesias, which is important to evaluate medication and surgical treatment.

This overview of methods for the objective assessment of LID illustrates that a successful method is not yet available. One of the reasons for this may be related to the limited set of tasks in which patients have been tested. The currently available algorithms for the assessment of LID were developed and applied to a small number of daily life activities, which were performed in a laboratory setting, each for a short duration (for example 1 minute). It may be that the data, collected in the small number of activities in these studies, did not contain enough information to provide an accurate measure to detect LID and to distinguish between LID and voluntary movements in daily life. If so, testing subjects over a longer period of time in a larger variety of activities might provide more and new information, which can be used by algorithms to detect and assess LID more accurately. Another improvement in classification performance may be obtained by recording movements in three orthogonal directions for various segments, because previous studies were limited to movements in two directions[13,14] or to measurement of movements of a single body segment.[15] In addition to an increased number of activities, testing patients in a natural environment for a long duration might provide more reliable data. A longer duration of testing will have the added advantage of showing various changing degrees of LID severity during the tests, providing insight into the movement variables which allow a distinction between LID and voluntary movements.

We tested patients with Parkinson's disease with various degrees of LID in a large variety of daily life activities for a period of a few hours in a natural environment to detect and assess the severity of LID. For the analysis of the data, we used neural networks that are known as adaptive techniques for complex classification problems and which can also provide valuable information on the movement variables that underlie a possibly successful detection and rating of LID.

PATIENTS AND METHODS

Patients

Thirteen patients with Parkinson's disease (8 men and 5 women) between 48 and 71 years old (mean, 61 ± 8 years) participated in this study. The patients had a mean duration of the disease of 15 ± 4 years (range, 10–21 years) and were on levodopa medication for several years. All patients suffered from LID. Mean levodopa medication was 692 ± 282 mg daily (range, 375–1,375 mg/day) and pergolide medication was 2.2 ± 2.5 mg daily (range, 0–8 mg/day). During the test, all patients showed a variety of grades of severity of LID. Seven patients showed a severity of dyskinesia varying between no dyskinesia and mild dyskinesia (rating between 0 and 1 on the AIMS scale). The other 6 patients showed a severity of dyskinesia varying between no dyskinesia to moderate (rating between 0 and 3 on the AIMS scale). The experiments were approved by the Medical Ethical Committee of the University Medical Center of the University of Nijmegen. The study started between 1200 and 1300 hours. The patients were continuously monitored for a period of approximately 2.5 hours. During this period, the patients took their regular medication at their usual time. However, when dyskinesia did not occur at the halfway point, extra levodopa was taken to induce dyskinesia.

The registration took place in a natural, home-like setting in the occupational therapy department of the University Medical Center. During the 2.5-hour monitoring session, the patients performed approximately 35 functional daily life activities, such as walking, putting on a coat, making coffee, preparing lunch, eating, taking off their shoes, reading a newspaper, drinking coffee, and washing hands. The order of the activities was randomized between subjects by a dedicated computer program. Subjects were allowed to carry out the activities in their own way and at their own pace. They were free to take a rest between activities at any time.

Data Acquisition

The movements and postures were automatically measured using accelerometers and a portable data recorder. Six sets of three orthogonal accelerometers (ADXL-202; Analog Devices, Norwood, MA) were used, which were placed at six different positions of the body. These six positions were at both upper arms (just below the shoulder), both upper legs (halfway the upper leg), at the wrist

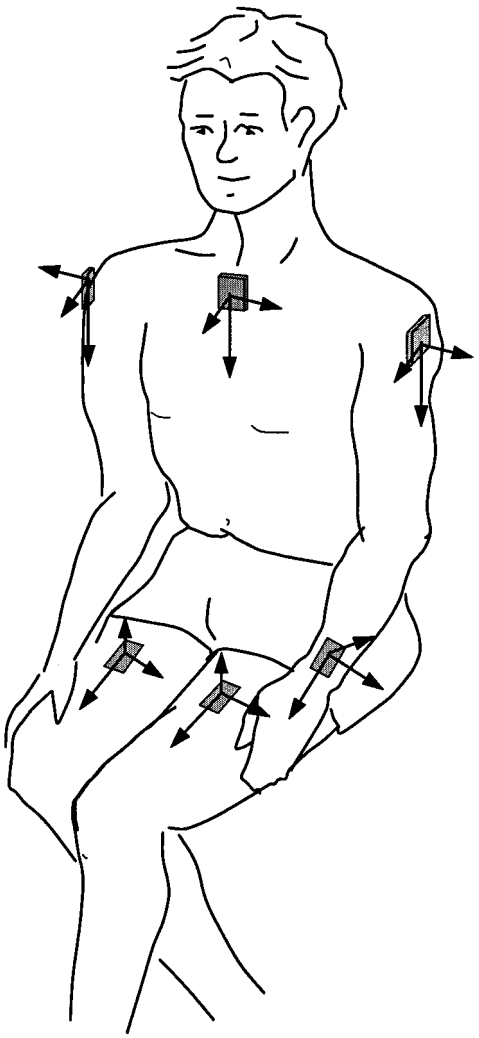

FIG. 1. Schematic overview of the position of accelerometers on the body. The directions for measurement of acceleration by each set of accelerometers are indicated by arrows.

of the most dyskinetic side, and at the trunk (top of the sternum; see Fig. 1). The accelerometer signals were digitally stored on a recorder (Vitaport 3; TEMEC Instruments, Kerkrade, The Netherlands) that was attached to a belt around the patient's waist. The accelerometer signals were sampled at a frequency of 256 Hz, low-pass filtered using a moving averaging window and stored at a sample frequency of 64 Hz. Advantages of this procedure are that it does not require a neurologist and that it easily can be placed within 15 minutes.

Thus far, the most reliable method to assess the severity of LID in daily life is to have the performance rated by experienced physicians. Therefore, the behavior of the patients was videotaped. The videotapes were used to rate the severity of LID on the modified AIMS scale[16] (m-AIMS) off-line by 2 experienced physicians, independently. The m-AIMS rating scale is a five-point scale with a value between 0 (absence of dyskinesia) and 4 (extreme dyskinesia).[16] Rating was done for each of the four limbs and for the trunk, separately. Data in a hypokinetic *off*-period without LID was excluded from further analysis.

Each start and end of an activity was stored on the data recorder using a radiographic system. A receiver was connected to the data recorder, and a sender was attached to a portable computer. When the patient started an activity, the experimenter pressed a key on the portable computer indicating the task that was started. The computer immediately transmitted a code to the receiver and the code was written on a separate channel of the data recorder worn by the patient. Simultaneously with recording onset and offset, an LED attached to the receiver was switched on and off. This switching LED informed the physicians to start or to end the video rating of LID.

Because different tasks had a different duration and because the severity of LID could fluctuate during an activity, we divided each task in subsequent time intervals of 1 minute, because a time resolution of 1 minute is clinically relevant and sufficient. Each 1-minute interval was evaluated separately, i.e., the severity of LID was video-rated by the physicians and the accelerometer characteristics were calculated for all subsequent 1-minute intervals.

Data Analysis

For each 1-minute interval signal, several variables were calculated from the accelerometer signals before being presented to the neural network. The neural network was trained with these variables as input and the rating scores given by the physicians as output. First, the preprocessing of the 1-minute accelerometer signals will be described, followed by the training and classification procedure with the neural network.

Preprocessing Accelerometer Signals.

Each raw accelerometer signal was filtered by a second-order low-pass digital Butterworth filter with a 3-dB cut-off frequency of 8 Hz. Accelerometers measure a contribution of gravity related to the orientation of the accelerometer and a contribution related to linear acceleration of the accelerometer. These components cannot be distinguished from each other. However, when there is movement, both components will change; thus, any change in the accelerometer signal will reflect movement of the accelerometer. For this reason, the derivative of the accelerometer signal was used as a measure of the amount of movement made by the subject. At each of the

TABLE 1. *Variables and their descriptions*

Symbol	Description
$\bar{V}$ segment	Mean segment velocity
$\bar{V}_{<3Hz}$ segment	The mean segment velocity for frequencies below 3 Hz
$\bar{V}_{>3Hz}$ segment	The mean segment velocity for frequencies above 3 Hz
$\frac{\bar{V}_{<3Hz}}{\bar{V}_{>3Hz}}$ segment	The ratio between $\bar{V}_{<3Hz}$ segment and $\bar{V}_{>3Hz}$ segment
SD(V) segment	The standard deviation of the segment velocity
% V_θ segment	Percentage of time that a segment was moving. A segment was considered as moving when the low-pass filtered segment velocity was above a threshold of about 0.05 m/sec
$\bar{V}_\theta$ segment	The mean segment velocity when the segment was considered to be moving, ie, when V segment $> V_\theta$ segment
P_{1-3Hz} segment	Power for frequencies in the range between 1 and 3 Hz
$P_{>3Hz}$ segment	Power for frequencies above 3 Hz
$\bar{\rho}_{segment-segment}$	The mean value of the normalized cross-correlation between the segment velocities of different segments
$\max(\rho_{segment-segment})$	The maximum value of the normalized cross-correlation between the segment velocities of different segments
% sitting	The percentage of time that a patient was sitting
% upright	The percentage of time that a patient's body was upright

Definition of the input variables to the neural network. The variables were calculated for each one-minute interval. The segment could be the most dyskinetic leg (mleg), the less dyskinetic leg (lleg), the most dyskinetic arm (marm), the less dyskinetic arm (larm) and the trunk (trunk). (For detailed explanation of the variables, see text.)

six body segments, we attached three accelerometers orthogonal to each other. To calculate the frequency and amplitude of body segment movements, we took the square root of the sum of squares of the derivatives of the three accelerometer signals from that body segment. The result will be referred to as "segment velocity."

For each of the body segments, the segment velocity was used to compute various variables for each 1-minute interval. The variables and their descriptions are shown in Table 1 and were calculated by a dedicated computer program. The first nine variables were calculated for each of the six different body segments. The variables *$\bar{V}$ segment, SD(V) segment, $\%V_\theta$ segment, and $\bar{V}_\theta$ segment* represent the mean velocity of a segment, the standard deviation relative to the mean velocity, the percentage of time a segment is moving, and the mean velocity when a segment moves, respectively. The variables *$\bar{V}_{<3Hz}$ segment, $\bar{V}_{>3Hz}$ segment, $\bar{V}_{<3Hz}/\bar{V}_{>3Hz}$* segment represent the mean segment velocity for frequencies below and above 3 Hz, and their ratio, respectively. These variables were used because it has been shown before that dyskinesia becomes manifest in the higher frequency domain.[13,15] Because the signal power for frequencies in the range between 1 and 3 Hz (*P_{1-3Hz} segment*) and above 3 Hz (*$P_{>3Hz}$ segment*) gave a good performance in classifying the severity of LID in the study of Manson and colleauges,[15] these accelerometer characteristics were also calculated. The cross-correlation between accelerometer signals from different body segments gives an indication of the coordination of movements of these segments. A high correlation (near one) indicates that movements of the two limb segments always covary, whereas a value near zero indicates that movements of the two limbs are uncorrelated. For this study, we calculated the mean cross-correlation between the velocity of two segments (*$\bar{\rho}_{segment-segment}$*) and the maximum of the cross-correlation (*$\max\rho_{segment-segment}$*). The percentage of the time a patient was sitting (*%sitting*) and/or when the patient's body was upright (*%upright*) were also used as variables. These variables were calculated using the accelerometer signals of the trunk and the leg in a similar way as in Veltink and coworkers.[17] The first nine variables were calculated for each of the six segments, which gave 54 different variables. Other variables were the mean value of the auto- and cross-correlation (n = 21) and the maximum value of the cross-correlation between movements of the six body segments gave another 36 variables. These variables, together with the percentage of time while the patient is sitting or while the patient's body was upright added another two variables, which brings the total number of variables to 92. All these variables were presented as input variables for the neural network.

Neural Network.

The neural network used in this study was a multilayer perceptron (MLP) with an input layer, one hidden layer, and an output layer. Each layer has several units and each unit is connected to all units in the next layer. As input variables, we used the variables derived from the accelerometer signals (see Table 1). The number of units in the hidden layer is crucial for the ability of the network to generalize, which is the ability to give a proper classification for a new input pattern, which the network has

not encountered before. There was one output unit for each body segment, the value of which reflects the severity of LID of that body segment. This segment could be the most dyskinetic arm, the trunk, or the most dyskinetic leg. The output of the units in the hidden layer was given by a hyperbolic tangent sigmoid transfer function that gives a value between -1 and $+1$. The output of the unit in the output layer was given by a linear transfer function and had a value in the range between 0 and 4, reflecting the AIMS score. Neural networks need a set of data, which provide examples of how sets of input values are related to the output (training set). The neural network uses these examples to adjust the weights between units in subsequent layers to minimize the error between the desired network output and the neural network output for each example. This is called a training process. After training, the network was tested using data, which were not used during the training process (test set). The neural network was trained using backpropagation. (For a review of neural networks, see Herz et al.[18])

Evaluating the Neural Network.

The performance of the network was evaluated using the mean square error (MSE) between the neural network output and the score given by the physicians. Because physicians could disagree in their rating, the mean value of the scores of the 2 physicians was used for training and testing the neural network. The physicians never had a difference in score larger than 1. In addition, the percentage of correctly classified signals by the neural network was used as a second criterion to evaluate the performance of the network. Because physicians rate dyskinesia by integers in the range between zero and four, the neural network classification was seen as correct when the difference between the neural network output and the score given by the physicians was smaller than 0.5. In other words, a classification was seen as correct when the rounded neural network output was exactly the same as that by the physicians.

The complexity of a network depends upon the number of units in the hidden layer and the number of variables used as input. A complex network will result in a good performance on a training set but can give a poor performance on a test set as a result of overfitting of the data set, i.e., the network has a poor generalization performance. For this reason, neural networks with various numbers of hidden units were trained to assess the severity of the most dyskinetic leg, the most dyskinetic arm, and the trunk. For each number of hidden units, the procedure of forward selection[19] was used to find the most valuable input variables to the neural network to assess the severity of LID. Forward selection means that we started with an empty variable set, and add, one after another, the variable that causes the largest reduction of the MSE between the neural network output and the score given by the physicians. After each step, we look for the next most important variable, and so forth. This procedure provides insight into the variables that are used by the neural network and that characterize its performance.

The generalization performance of the network was tested by training the network with 80% of the data set and testing the network with the remaining 20% of the data. This procedure was done 50 times for different randomly selected sets of training and test sets. The optimal architecture of the network was seen as the network, which gave on average the smallest MSE on the test set for the 50 randomly selected sets.

The first goal of the study was to test the possibility of detecting and assessing the severity of LID for patients with Parkinson's disease by studying a large variety of daily life activities, i.e., the network's ability to generalize over various tasks. However, the network should also be able to classify the severity of LID for new patients, which the network has never seen before, i.e., the network should also be able to generalize over patients. The network architectures with the best performance in detecting and assessing the severity of LID in a large variety of daily life activities were used to test the performance for new patients. For this testing, the neural network was trained with all data except for the data of one patient ("leave one patient out"). The data of the remaining patient was predicted using the trained neural network. This "leave-one-out" method was applied for each patient and gave a good impression of the ability of the network to classify the severity of LID for patients, which the network has not seen before. The performance of the network was evaluated using two measures: the MSE between the neural network output and the score given by the physicians, and the percentage of correctly classified data.

RESULTS

Figure 2 shows the MSE for the training set (open symbols and dashed lines) and test set (filled symbols and solid lines) for neural network architectures with various numbers of units in the hidden layer. The MSE is plotted as a function of the number of input variables for the most dyskinetic leg ordered according to their relevance for the detection and assessment of LID. As shown in Figure 2, the MSE starts to decrease when the number of input variables increases for each number of hidden units. When the number of input variables becomes

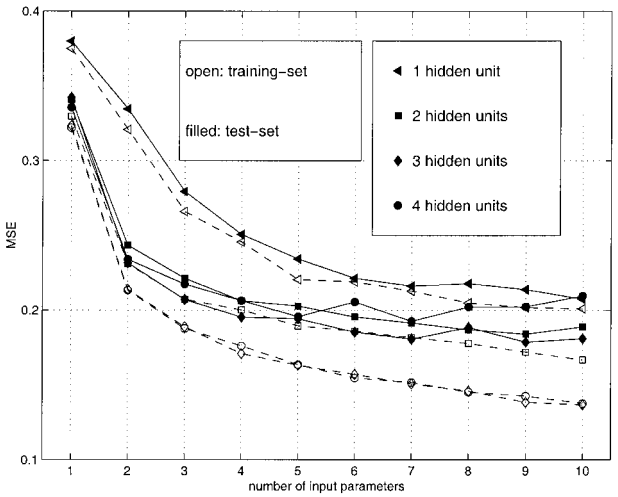

FIG. 2. The mean square error on the training set (open symbols and dashed lines) and test set (filled symbols and solid lines) for neural networks with 1, 2, 3, and 4 units in the hidden layer. Each network was trained with data for the most affected leg using the input parameters (see *Patients and Methods* and Table 1) and the rating by the physician.

larger than four, the MSE for the training set decreases only slightly. The MSE on the test set shows initially a decrease for each added variable, followed by an increase in MSE when the number of input variables becomes larger. The increase in MSE on the test set for large numbers of input variables is the result of overfitting of the data.

The network with three hidden units and seven input variables as input gave the best performance on the test set (smallest MSE) for the most dyskinetic leg (see Fig. 2). For the arm, a network with two hidden units and six variables as input gave the smallest MSE on the test set. For the trunk, the best performance was obtained for a network with only one hidden unit. The optimal number of input variables appeared to be relatively large (n = 12). Table 2 shows a list of the relevant input variables, which result from the neural network and the forward selection procedure, in order of importance for each of the three segments. A variety of variables are important and the important variables differ for different body segments. For the arm and especially for the trunk, variables related to movements of other body segments appeared to be relevant. For the leg, variables of both legs and the trunk and the cross-correlation between these segments appeared to be relevant.

The MSE and the percentage of correctly classified data on the training and test sets for the best performing networks on the test set for 1-minute intervals are shown in Table 3. The results in Table 3 indicate that, in general, the error between the score by the physicians and by the neural network (0.19 or less) is small relative to the AIMS scale, which ranges between 0 and 4 with integer increments. The percentage of correctly classified 1-minute intervals on the test set has the largest value for the trunk (83.0 ± 4.0%) and was slightly smaller for the arm (77.0 ± 3.1%) and the leg (76.9 ± 3.9%). The correlation coefficients between the neural network output on the test set and the physicians rating were 0.71, 0.87, and 0.80 for the arm, trunk, and leg, respectively.

Figure 3 shows an example of the scores given by the physician and the scores given by the neural network on a test set for 81 one-minute intervals. These 81 one-minute intervals were taken out of the 2.5-hour session of a patient in which periods of rest were not shown to present the performance for a representative set of activities. The scores predicted by the neural network do agree well with the scores given by the physicians. Both scores change almost simultaneously in time over the time interval of 81 minutes. For scores for which the physicians disagree (in these cases, the average of the physician's score was 0.5, 1.5, or 2.5), the network gave a value between the scores given by the physicians. In general, the difference in rating given by the physicians and the network is 0.5 or lower. Because the patient showed only mild symptoms of dyskinesia, Figure 3 shows that the neural network was sensitive and accurate in detecting LID.

The neural network classification was considered to be correct when the difference between the neural network output and the score given by the physicians was lower

TABLE 2. *Relevant input variables*

Stage	Arm	Trunk	Leg
1	$\frac{\bar{V}_{<3Hz}}{\bar{V}_{>3Hz}}$ mleg	% V_θ trunk	SD(V) lleg
2	$\bar{\rho}_{wrist-trunk}$	SD(V) lleg	% V_θ mleg
3	% V_θ mleg	$\bar{V}_{<3Hz}$ Trunk	% sitting
4	% V_θ wrist	% V_θ lleg	$\bar{\rho}_{lleg-trunk}$
5	$\bar{\rho}_{wrist-larm}$	$P_{>3Hz}$ marm	P_{1-3Hz} trunk
6	% sitting	P_{1-3Hz} mleg	$\bar{V}_\theta$ mleg
7	—	$\bar{\rho}_{mleg—trunk}$	$\max(\rho_{mleg-trunk})$
8	—	$P_{>3Hz}$ mleg	—
9	—	P_{1-3Hz} lleg	—
10	—	$\bar{V}_{>3Hz}$ marm	—
11	—	$\frac{\bar{V}_{<3Hz}}{V_{>3Hz}}$ larm	—
12	—	SD(V) wrist	—

Input variables, relevant for the detection and classification of LID for the arm, trunk and leg, in order of importance. The order of importance was determined using forward selection for the network with the smallest mean square error (MSE) between neural network output and the score given by the physicians on the test set for the arm (2 hidden units), the trunk (1 hidden unit) and the leg (3 hidden units). Subscripts refer to marm (most dyskinetic arm); larm (less dyskinetic arm); mleg (most dyskinetic leg) and lleg (less dyskinetic leg).

TABLE 3. *Data from the best performing networks*

	MSE (1-min interval)		% good (1-min interval)		
Segment	Training set	Test set	Training set	Test set	% good (15 min)
Arm	0.17 ± 0.01	0.19 ± 0.02	78.3 ± 0.9	77.0 ± 3.1	93.7
Trunk	0.14 ± 0.01	0.14 ± 0.02	83.4 ± 0.9	83.0 ± 3.4	99.7
Leg	0.15 ± 0.01	0.18 ± 0.03	80.5 ± 1.4	76.9 ± 3.9	97.0

Performance of the neural network averaged over all one-minute time intervals of the 2.5 hours session (columns 2, 3, 4, and 5) and the percentage of correctly classified data in 15-minute time interval (column 6). Performance of the neural network is expressed by the mean and standard deviation of the mean square error (MSE) between the neural network output and the score given by the physicians (columns 2 and 3) and by the percentage of correctly classified activities (% good) (columns 4, 5 and 6) for the arm, trunk and leg.

MSE, mean square error.

than 0.5. It would be interesting to see what the percentage of correctly classified data would be for other error margins between the neural network output and the physician's score. Figure 4 shows the percentage of correctly classified 1-minute intervals on the test set as a function of the error margin for the arm, leg, and trunk for the whole population of data. More than 95% of the 1-minute intervals had a difference less than 0.85 between neural network output and the score given by the physicians. When differences up to 1.0 were allowed, more than 98.0% of the 1-minute intervals were classified correctly. This finding suggests that, if the rating by the neural network were different from the rating given by the physician, it was in the grade next to the score given by the physicians.

From a clinical point of view, physicians are mainly interested in whether patients suffer from dyskinesia for at least a few minutes. Therefore, we determined the performance of the network under the constraint that it should correctly predict dyskinesia or the absence of dyskinesia for longer periods. For periods of 15 minutes, the neural network correctly classified dyskinesia or absence of dyskinesia in 93.7, 99.7, and 97.0% for the arm, trunk, and leg, respectively (see Table 3). The correlation coefficient between the neural network and the physicians rating averaged over 15 minutes were 0.88, 0.96, and 0.92 for the arm, trunk, and leg, respectively.

Recently, Manson et al.[15] reported a good Spearman rank correlation between acceleration signals in the 1 to 3 Hz frequency band and the rating on the modified AIMS scale. For the data in our study, the Spearman rank correlation between the acceleration in the 1 to 3 Hz frequency (P_{1-3Hz} *segment*) and the m-AIMS score for the arm, trunk, and leg was 0.18, 0.30, and 0.21, respectively. As shown in Table 2, the neural network indicated other variables with more predictive power in addition to

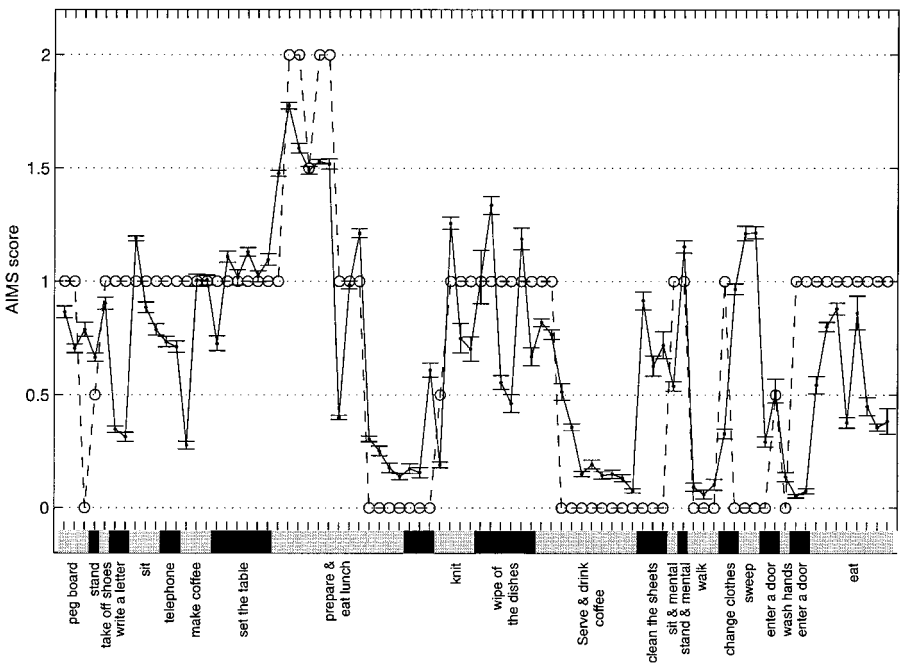

FIG. 3. Example of the AIMS rating given by the physicians (circles) and predicted by the neural network (dots with error bars) for the trunk for 81 one-minute intervals of various activities.

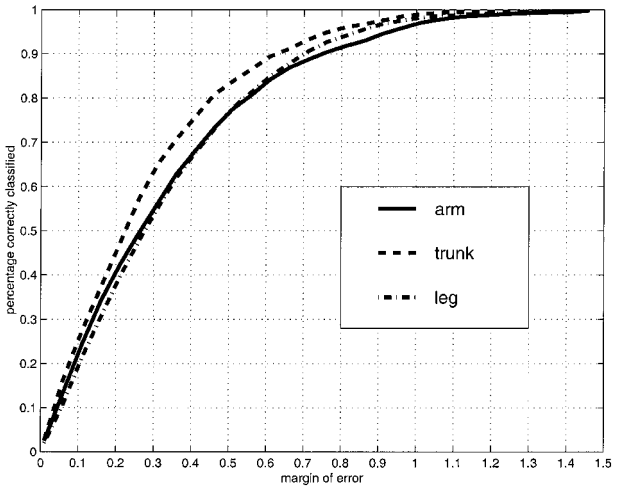

FIG. 4. Percentage of correctly classified data in the test set as a function of the error margin for the arm (solid line), trunk (dashed lines), and leg (dashed-dotted line).

the acceleration in the 1 to 3 Hz frequency band. The most valuable variables, the ratio between low and high frequencies of the most affected leg, the percentage of the time that the trunk was moving, and the standard deviation of the leg (see Table 2), gave a Spearman rank correlation of 0.38, 0.44, and 0.37 for the arm, trunk, and leg, respectively. This finding means that these most valuable variables contribute approximately 2 to 4 times more in explaining the AIMS score than the acceleration in the 1 to 3 Hz frequency range for the data in our study.

To demonstrate the neural network's ability to distinguish LID from voluntary movements, the performance of the network was evaluated for three different groups of activities. The first group included the activities sitting and standing with or without a mental task. During these activities, patients were ordered to abstain from any voluntary movement and not to suppress any involuntary movements. The second group consisted of activities for which patients now and then made voluntary movements. This second group included activities such as drinking coffee, reading a newspaper, making a phone call, and writing. The third group consisted of activities for which patients made voluntary movements for almost the entire period such as making coffee, walking, setting the table, dressing, etc. The performance of the neural network output was considered to be correct when the neural network gave a value smaller than 0.5 for the 1-minute intervals, which were rated by the physicians with the score 0 (no dyskinesia group), and when the network gave a score larger than 0.5 for the 1-minute intervals, which were rated by the physicians with a rating 1 or higher (dyskinesia group). The percentage of correctly classified minutes for the different groups is shown in Table 4. The correct performance of the neural network is between 75% and 100%, depending on the type of movements. The best performance is obtained in the absence of voluntary movements and in the absence of dyskinesia. The network displayed some tendency to erroneously detect absence of dyskinesia in patients with mild dyskinesia who were trying to abstain from any voluntary movements. This is primarily because normal subjects, when sitting in a relaxed position, make small movements with the legs and arms that are hard to distinguish from mild dyskinesia. In general, the neural network was able to correctly distinguish the large majority of LID movements from voluntary movements.

A more detailed overview of the rating performance by the neural network for various types of behavior with voluntary movements is shown in Figure 5, which shows the percentage of correctly classified behavior for a selection of activities. In general, approximately 80% of the 1-minute intervals of each activity was correctly classified. Classification algorithms in previous studies showed discrepancies with the rating by physicians for activities with voluntary movements and especially for walking.[13–15] The neural network gave an extremely well-fit classification for 1-minute intervals of walking for the trunk (100%) and the leg (96%), but less so for the arm (61%).

TABLE 4. *Correctly classified minutes*

	Segment	No voluntary movements	Now-and-then voluntary movements	Many voluntary movements
Absence of dyskinesia (AIMS = 0)	Arm	100.0	79.6	78.5
	Trunk	100.0	98.5	88.3
	Leg	92.6	80.2	77.2
Dyskinesia (AIMS ≥ 1)	Arm	75.0	90.0	79.4
	Trunk	94.6	84.7	90.4
	Leg	76.9	87.3	82.6

Percentage of correctly classified one-minute intervals with dyskinesia and absence of dyskinesia for time intervals without voluntary movement, intervals with activities requiring voluntary movements some now and then, and for intervals with activities which require frequent voluntary movements.

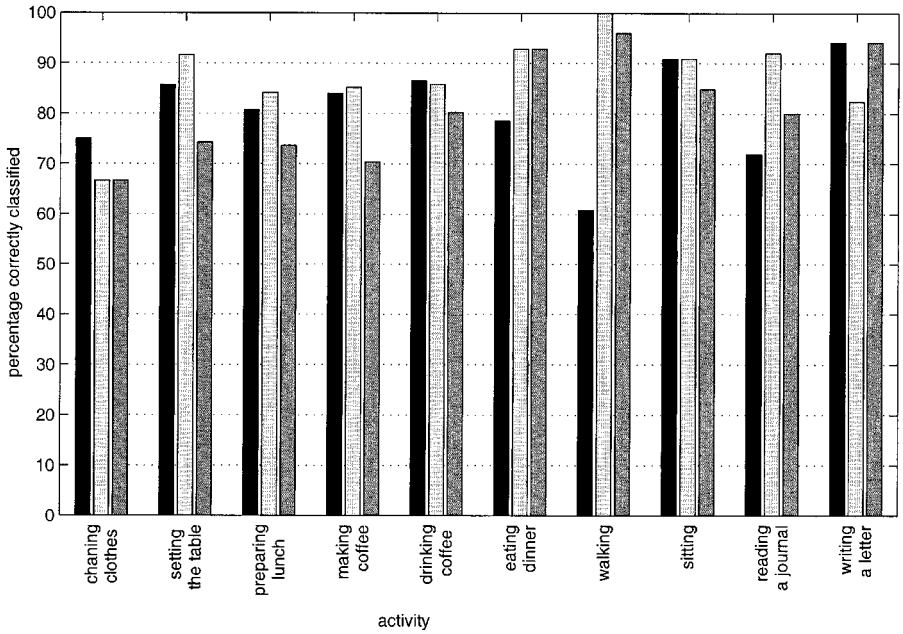

FIG. 5. Percentage of correctly classified data for various activities. For each grouping, the first bar is for the arm, second bar is for the trunk, and the third bar is for the leg.

To test the ability of the network to classify the severity of LID for patients that the network has not seen before, neural networks were trained with all data except for the data of 1 patient. Thereafter, the trained network was used to predict the severity of LID for the remaining patient. The mean and standard deviation of the MSE and the percentage of correctly classified 1-minute intervals for the various subjects for the arm, trunk, and leg are shown in Table 5. The performance of the network for data in a 15-minute interval is also shown in Table 5 (see column 6). The performance of the networks is approximately the same as that shown in Table 3, indicating that the neural network could equally well generalize over activities and subjects.

DISCUSSION

Recent studies have indicated the validity of ambulatory accelerometry in assessing the severity of LID.[12–15] However, previous studies have not been adequately sensitive, nor can they distinguish between voluntary movements and LID. Another important limitation of studies to date has been the small number of tasks involved, and the fact that they have been performed in a laboratory setting. In the present study, patients performed a large variety of daily life activities in a natural environment for a long duration. The neural network was able to detect and assess the severity of LID correctly for a large fraction of tasks. When the rating by the neural network differed from the rating given by the physicians, the difference in rating was small, and in the worst cases the rating was in the grade next to that indicated by the physicians.

Previous studies[12,13,15] have used linear classification techniques to detect and to assess LID. In our previous study,[14] the best performing neural network did have one hidden unit, which is equivalent to a linear classification. With the larger and richer data set in this study, we found that the optimal number of hidden units for the neural network for rating LID is three for the leg and two for the

TABLE 5. *Data from the network for 'leave one patient out'*

	MSE (1-min interval)		% good (1-min interval)		
Segment	Training set	Test set	Training set	Test set	% good (15 min)
Arm	0.17 ± 0.01	0.22 ± 0.10	78.4 ± 1.1	74.0 ± 11.8	93.6 ± 15.1
Trunk	0.14 ± 0.01	0.15 ± 0.11	83.3 ± 0.9	82.4 ± 16.5	99.5 ± 1.7
Leg	0.15 ± 0.00	0.20 ± 0.08	81.2 ± 0.8	70.3 ± 14.7	92.7 ± 11.1

Performance of the neural network over different patients using the leave-one-output method. Performance of the neural network (mean square error (MSE) and the percentage correctly classified activities by the neural network (% good) for one-minute intervals (columns 2, 3, 4, and 5). The percentage of correctly classified data for 15-minute time interval data (column 6).

MSE, mean square error.

arm. This finding indicates that nonlinear interactions between various movement variables (which may not have been obvious in our previous study due to the limited number [n = 7] of tasks) are important for the proper rating of LID. For the trunk, the best performing neural network had only one unit in the hidden layer, indicating that a linear technique may be sufficient. However, in this case, the number of input variables, which contribute information to the detection and classification of LID, appeared to be relatively large (n = 12).

In our comparison of rating by physicians and by the neural network, we have used the averaged rating by the physicians. Assuming that experienced physicians rate LID in the same class or in neighboring classes, we considered the rating by the neural network incorrect when the rating by the neural network differed by more than 0.5 from the average of the rating by the physicians. With that criterion, approximately 80% of the 1-minute intervals were classified correctly (see Tables 3 and 5). However, the criterion of 0.5 may be somewhat arbitrary. The physician's rating is semiquantitative and is not sensitive to small changes. Moreover, the rating by the physicians will presumably be affected by ratings in previous minutes. In many cases, we observed that the changes in the rating by the neural network anticipated those by the physicians. These influences on the physician's rating triggered us to consider the score for other error margins. When the error margin was extended to 1.0, the correct score went up to more than 98.0%. Irrespective of the question of which error margin to use, our results demonstrate that any differences between the rating by the physicians and the neural network do not differ by more than one grade on the AIMS scale.

Another aspect of the rating by the neural network was that any difference with the rating by the physicians usually lasted for 1 minute only. When periods with a longer duration where evaluated, the error rate decreased and the correct performance increased to 93.7, 99.7, and 97.0 % for the arm, trunk, and leg, respectively. These results indicate that the procedure described in this study to detect and assess LID seems a valid method for practical use.

A major advantage of using neural networks for the detection and rating of LID with the forward selection procedure to find the most relevant variables is that this procedure searches for the most valuable variables without any prior information and restriction. In general, the percentage of time that a segment was moving ($\%V_{\theta}$ *segment*), the cross-correlation between segments ($\bar{\rho}_{segment-segment}$), and variables evaluating the signals in the frequency domain appeared to be the most important variables. These variables are in line with the most important variables found in a previous studies.[13–15]

One of the most important variables appeared to be the percentage of time that the arm, trunk, or leg was moving. The importance of this variable is obvious, because a small percentage indicates few movements and probably no dyskinesia, whereas a large percentage indicates many movements and, thus, a possibility that the subject might suffer from dyskinesia.

One of the main difficulties in assessing LID is the ability to distinguish LID from voluntary movements. Hoff and colleagues[13] and Manson and associates[15] reported that acceleration signals in the 1 to 3 Hz frequency band correlated well with the modified AIMS scale and stated that dyskinesias occur in a higher frequency domain than voluntary movements. In our analysis, the acceleration signals in the range between 1 and 3 Hz also appeared to be a variable, which contributes to the detection and rating of LID. However, the power of the acceleration signals in the 1 to 3 Hz frequency domain explained only a small fraction of the severity of LID. This finding indicates that the frequency range of accelerometer signals of voluntary movements is not disjunct from that of the accelerometer signals for dyskinesias, which is in agreement with previous reports.[13,20]

Moreover, the neural network analysis revealed several other variables which can contribute to distinguish LID from voluntary movements, such as the cross-correlation between acceleration signals from two different limb segments and by comparing the movements of various limb segments. This can be understood from the fact that dyskinesia is frequently observed in multiple body segments.[21] In our study, this resulted in a small value of the correlation between movements of these body segments combined with high values for the percentage of time of moving for these body segments. We also observed that patients suffering from mild dyskinesias showed dyskinesia only in a single limb or in the trunk. In such cases the correlation coefficient was zero if one of the body segments did not move. In case of dyskinesia superimposed on voluntary movements, such as in walking, the correlation between movements of the arm and leg does not provide much information. In that case, the power in the frequency range below and above 3 Hz was used to detect dyskinesia, in agreement with the results of previous studies. A detailed description of the contribution of various parameters requires more sophisticated analyses, which is outside the scope of this study.

Our results showed that the neural network was able to distinguish LID from voluntary movements (see Table 4). The performance of the neural network was slightly less for the group of patients with dyskinesia, who ab-

stain from voluntary movements, and for the group of patients without dyskinesia, who made many voluntary movements. For the group of patients with dyskinesia who abstained from voluntary movements, the neural network had some difficulty to distinguish mild dyskinesia from normal small movements of the arm and the leg, which occur now and then when subjects sit relaxed for some time. In normal daily life, patients hardly ever abstain completely from any voluntary movements. Therefore, the second group of tasks, wherein patients occasionally made voluntary movements, may be more illustrative for daily life situations with few voluntary movements. For this second group of tasks, the neural network showed a good performance in detecting dyskinesia.

The network rated some voluntary movements as dyskinesia for patients with absence of dyskinesia who made many voluntary movements. This misclassification is most frequently observed in activities such as washing the dishes or sweeping the floor. These typical activities show voluntary movements that contain movement characteristics similar to that of dyskinesia.

The obvious question to ask is: what explains the better performance of rating in this study relative to that in previous studies? A possible explanation is that previous studies used a limited set of tasks, which had to be performed in a highly controlled laboratory setting.[12–15] This strategy may have resulted in a limited data set with possibly some unnatural behavior of the patients. The present study tested patients with varying degrees of severity of LID in a large variety of daily activities. This larger number of activities and varying degree of severity of LID provides more information for the adaptive neural networks to find the proper variables to distinguish between voluntary movements and LID. These variables and their mutual linear and nonlinear connections are probably not disclosed with the methods used by other investigators. The next step will be to investigate how the neural network combined the various variables for rating. This will provide more information about the characteristics of LID in comparison to that of voluntary movements.

In conclusion, our method accurately assessed the severity of LID and distinguished LID from voluntary movements in a daily life situation. The difference between the neural network output and the score by the physicians was small and, worst case, the rating by neural networks was in the class next to that indicated by the physician. Therefore, the method used in this study could be operating successfully in unsupervised ambulatory conditions.

Acknowledgments: This study was financially supported by the Prinses Beatrix Fonds (MAR 00-104) and the Parkinson Patienten Vereniging. We thank the Department of Occupational Therapy of the University Medical Center St. Radboud for the opportunity to use their facilities for the measurements in this study. We also thank Hans Kleijnen for his technical support and Linda Mol and Co van de Lee for their contribution to this article.

REFERENCES

1. Marsden CD. Parkinson's disease. J Neurol Neurosurg Psychiatry 1994;57:672–681.
2. Nutt JG, Carter RN, Woodward WR. Long-duration response to levodopa. Neurology 1995;45:1613–1616.
3. Brotchie J. Adjuncts to dopamine replacement: a pragmatic approach to dealing with the problem of dyskinesia in Parkinson's disease. Mov Disord 1998;13:871–876.
4. Manson AJ, Schrag A, Lees AJ. Low-dose olanzapine for levodopa induced dyskinesias. Neurology 2000;55:795–799.
5. Fraix V, Pollak P, Van Blercom N, Xie J, Krack P, Koudsie A, Benabid AL. Effect of subthalamic nucleus stimulation on levodopa-induced dyskinesia in Parkinson's disease. Neurology 2000; 55:1921–1923.
6. Goetz CG. Rating scales for dyskinesias in Parkinson's disease. Mov Disord 1999;14(Suppl. 1):48–53.
7. Damier P, Jaillon C, Clavier I, Arnulf I, Bonnet A, Bejjani B, Agid Y. Dyskinesia assessment in phase II studies. 1999;14(Suppl. 1):54–59.
8. Widner H, Defer G. Dyskinesia assessment: From CAPIT to CAPSIT. 1999;14(Suppl. 1):60–66.
9. Golbe LI, Pae J. Validity of a mailed epidemiological questionnaire and physical self-assessment in Parkinson's disease. Mov Disord 1988;3:245–254.
10. Vitale C, Pellecchia MT, Grossi D, Fragassi N, Cuomo T, Di Maio L, Barone P. Unawareness of dyskinesia in Parkinson's and Huntington's diseases. Neurol Sci 2001;22:105–106.
11. Brown P, Manson A. Dyskinesia assessment and ambulantory devices. 1999;14(Suppl. 1):67–68.
12. Burkhard PR, Shale H, Langston W, Tetrud JW. Quantification of dyskinesia in Parkinson's disease: validation of a novel instrumental method. Mov Disord 1999;14:754–763.
13. Hoff JI, van de Plas AA, Wagemans EAH, van Hilten JJ. Accelerometric assessment of levodopa-induced dyskinesias in Parkinson's disease. Mov Disord 2001;16:58–61.
14. Keijsers NLW, Horstink MWIM, van Hilten JJ, Hoff JI, Gielen CCAM. Detection and assessment of the severity of levodopa-induced dyskinesia in patients with Parkinson's disease by neural networks. Mov Disord 2000;15:1104–1111.
15. Manson AJ, Brown P, O'Sullivan JD, Asselman P, Buckwell D, Lees AJ. An ambulatory dyskinesia monitor. J Neurol Neurosurg Psychiatry 2000;68:196–201.
16. Guy W, editor. AIMS. In ECDEU assessment manual. Rockville, MD: US Department of Health, Education, and Welfare; 1976. p 534–537.
17. Veltink PH, Bussmann HBJ, de Vries W, Martens WLJ, van Lummel RC. Detection of static and dynamic activities using uniaxial accelerometers. IEEE Rehab Eng 1996;4:375–385.
18. Hertz J, Krogh A, Palmer RG. Introduction to the theory of neural computation. Redwood City, CA: Addison-Wesley; 1991.
19. Laar P, Heskens TM, Gielen CCAM. Partial retraining: a new approach to input relevance determination. Int J Neural Syst 1999; 9:75–85.
20. Redmond DP, Hegge FW. Observations on the design and specification of a wrist-worn human activity monitoring system. Behav Res Methods Instrum Comput 1985:659–669.
21. Marconi R, Lefebre-Caparros D, Bonnet AM, Vidailhet M, Dubois B, Agid Y. Levodopa-induced dyskinesias in Parkinson's disease; phenomenology and pathophysiology. Mov Disord 1994;9:2–12.