Clinimetrics, clinical profile and prognosis in early Parkinson's disease
Post, B.

Citation for published version (APA):

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: http://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.
Chapter 2

Unified Parkinson’s disease rating scale motor examination: are ratings of nurses, residents in neurology, and movement disorders specialist interchangeable?

B. Post
M.P. Merkus
R.M.A. de Bie
R.J. de Haan
J.D. Speelman

Movement disorders 2005; 20: 1577-1584
Abstract

Objective

The Unified Parkinson’s Disease Rating Scale (UPDRS) is widely used for the clinical evaluation of Parkinson’s disease (PD). We assessed the rater variability of the Unified Parkinson’s Disease Rating Scale motor examination (UPDRS-ME) of nurse practitioners, residents in neurology and a movement disorders specialist (MDS) compared to a senior MDS.

Methods

We assessed the videotaped UPDRS-ME of 50 PD patients. Inter-rater and intra-rater variability were estimated using weighted kappa ($\kappa_w$) and intra class correlation coefficients (ICC). Additionally, inter-rater agreement was quantified by calculation of the mean difference between two raters and its 95% limits of agreement. Intra-rater agreement was also estimated by calculation of a 95% repeatability limits.

Results

The $\kappa_w$ and ICC statistics indicated good to very good inter-rater and intra-rater reliability for the majority of individual UPDRS-items and the sum score of the UPDRS–ME in all raters. However for inter-rater agreement, it appeared that both nurses, residents, and the MDS consistently assigned higher scores than the senior MDS. Mean differences ranged between 1.7 and 5.4 (all differences P<0.05) with rather wide 95% limits of agreement. The intra-rater 95% repeatability limits were rather wide.

Conclusion

We found considerable rater difference for the whole range of UPDRS-ME scores between a senior MDS and nurse practitioners, residents in neurology, and a MDS. This implies that the amount by which raters may disagree should be quantified before starting longitudinal studies of disease progression or clinical trials. Finally, evaluation of rater agreement should always include the assessment of the extent of bias between different raters.
Introduction

The Unified Parkinson’s Disease Rating Scale (UPDRS) is widely used for the clinical evaluation of Parkinson’s disease (PD). (1) The scale consists of the following four components: part one mentation, behaviour and mood; part two activities of daily living; part three motor examination; and part four complications of therapy. (2) The motor examination part of the UPDRS (UPDRS-ME) quantifies type, number, and severity of extra-pyramidal signs. The UPDRS-ME has a high internal consistency (3-5) and is responsive to changes in time. (6) The inter-rater reliability of expert neurologists has been shown to be adequate. (3,7-10) The intra-rater reliability ranged from low to good. (8,10,11) Administration of the UPDRS-ME is generally believed to require experienced raters. (6,7,12) In search of the most effective and efficient combination of health care professionals to deliver care for chronic patients, nowadays nurse practitioners perform tasks that traditionally belonged to the domain of the physician. Regarding care of PD patients, this is reflected by the fact that the administration of the UPDRS-ME is more and more handed over to nurse practitioners. The ability of nurse practitioners to administer the UPDRS-ME has been reported once. (8) This study reported ratings of nurse practitioners corresponding closely with those of a neurologist specialised in movement disorders. However, this concerned a modified version of the UPDRS-ME that was applied to a community sample of elderly with mild parkinsonian symptoms who did not come to medical attention.

In addition, although residents in neurology are fully enlisted in the care of patients with PD, their ability to administer the UPDRS-ME has never been formally assessed. A previous study reported a satisfactory inter-rater reliability of the UPDRS-ME for residents in geriatrics. (3) Therefore, we assessed the inter-rater and intra-rater variability of the UPDRS-ME of nurse practitioners, residents in neurology and a movement disorders specialist (MDS) compared to a senior MDS.

Methods

Patients

Fifty patients with PD participating in research projects conducted at our clinic were included. Twenty PD patients were recruited from an ongoing prospective cohort study on functional status and had PD diagnosed after 1996. Thirty PD patients were derived from a Dutch multi-centre trial assessing the efficacy of unilateral pallidotomy. (13) All
the patients were assessed by a resident in neurology to confirm the diagnosis of PD according to internationally accepted diagnostic criteria.(14)

**UPDRS-ME**

The UPDRS-ME allows the scoring of speech, facial mobility, resting tremor (face and each limb), action tremor (upper limbs), finger taps (taps of thumb with index finger), hand movements (rapid opening and closing), rapid alternating movements (pro- supination), leg agility (rapid heel tapping), arising from a chair, posture, gait, postural stability and body bradykinesia / hypokinesia. For each sign a five-step severity gradation is used, with zero representing absence and four representing maximum severity of that sign.(2)

**Raters**

The raters were two nurse practitioners, two residents in neurology, and two MDS. Although both nurse practitioners already had experience scoring the UPDRS-ME, they underwent a formal training protocol before starting this study: first, the UPDRS-ME was explained by the senior MDS (J.D.S.); second, the teaching tape for UPDRS-ME developed by Goetz et al. (9) was studied; and third, 10 videotaped administrations of the UPDRS-ME of patients with PD were scored and discussed in the presence of all raters participating in the study.

**Procedure**

The UPDRS-ME was administered after a clinical evaluation, containing a history and a neurological examination. Patients were taped according to a standard protocol, which was based upon the teaching tape for the UPDRS-ME.(9) All subjects gave informed consent. Rigidity was not recorded because it cannot be assessed on videotape.(9,15) The patients from the trial were all scored in the standardised off-state (13), because in the on-state they all had dyskinesias, which interfere with measurement of the UPDRS-ME.(9) The patients from the cohort study were assessed in the outpatient clinic. They were all on medication and only one of them had mild dyskinesias. All motor signs were rated using the videotape. The raters did not discuss their ratings with each other during or after the assessment of the patient, and thus remained unaware of each others scores. For the assessment of the inter-observer reliability all videotaped patients were rated by the two nurse practitioners, two residents in neurology (BP, AWL) and one MDS (RdB). All ratings were compared to the ratings of the senior MDS (JDS). One nurse practitioner, one resident in neurology (BP) and the senior MDS (JDS) rated all patients using the same videotape for a second time 6-8 weeks later to assess intra-rater reliability.
Data analysis

The inter-rater variability was calculated for each item using a weighted kappa-statistic ($\kappa_w$) for each pair of observers. The kappa statistic estimates the proportion of agreement among observers after chance agreement has been removed. Chance agreement is the proportion of agreement expected if observer’s ratings were completely random. A $\kappa_w$ adds different weights to disagreement according to the magnitude of discrepancy in ordinal data. If kappa equals zero it denotes that the agreement between the examiners is not better than that expected by chance and if the kappa equals one there is perfect agreement. There is no absolute definition how to interpret kappa values, but we used criteria proposed by Altman (17), and adapted from Landis and Koch 1977 (18); <0.20 poor agreement ; 0.21-0.40 fair agreement; 0.41-0.60 moderate agreement; 0.61-0.80 good agreement; and 0.81-1.0 very good agreement. For the total score of the UPDRS motor section inter-rater reliability was evaluated with help of the intraclass correlation coefficient (ICC). The ICC is a parametric measure assessing rating reliability by comparing the variability between subjects to the total variability (subjects, raters, error). The values of the ICC range between 0 and 1, with a higher value indicating that less variance is due to other factors such as differences between observers. The ICC was calculated as a two-way mixed effects model with absolute agreement. Additionally inter-rater agreement was studied in a scatter plot as described by Bland and Altman. This plot shows the difference between the two measurements by each rater on the vertical axis against their mean on the horizontal axis. Vertical spread correlates with the degree of inter-rater variation for the specific value. To predict the possible range of differences, 95% limits of agreement were calculated as 1.96 times the standard deviation of the mean difference, provided the differences are normally distributed. This means that we would expect 95% of the differences between two raters to lie between these limits. For the interpretation of the measurement error, the largest limit of agreement (either upper or lower limit) is most relevant. To assess systematic bias the mean difference is used. In case of systematic bias the mean difference 95% confidence interval will not include zero. The strength of the relationship between the differences and their mean was indicated by the slope of a regression line. The intra-rater variability was calculated using the $\kappa_w$ and the ICC. The ICC for the intra-rater reliability was calculated using one way random effects model. Additionally, 95% repeatability coefficients (or intra-rater agreements) were calculated. STATXACT-version 3.0 statistical software was used to calculate the weighted kappa statistic. SPSS -version 11.5 statistical software was used for all other statistical analyses.
Results

Patient characteristics

Demographic and clinical characteristics of the PD patients are displayed in Table 1. Age ranged from 35 to 83 years with a mean age of 63.2 years (standard deviation (SD) 19.3 yrs). Mean disease duration amounted to 11.4 years (SD 7.4) with a range of 0.5-27.5 years. All Hoehn and Yahr stages were represented in the sample. The range of the sum scores of UPDRS-ME was 7 to 82 points (mean 37.7), so it encompassed the whole spectrum of PD.

<table>
<thead>
<tr>
<th>Clinical variables</th>
<th>Total (n=50)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Current age (mean yrs (SD, range))</td>
<td>63.2 (19.3; 35-83)</td>
</tr>
<tr>
<td>Age of onset (mean yrs (SD, range))</td>
<td>51.8 (12.4; 27-78)</td>
</tr>
<tr>
<td>Gender (m/f)</td>
<td>25/25</td>
</tr>
<tr>
<td>MHY-score</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td>1.5</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>9</td>
</tr>
<tr>
<td>2.5</td>
<td>5</td>
</tr>
<tr>
<td>3</td>
<td>13</td>
</tr>
<tr>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>5</td>
<td>12</td>
</tr>
<tr>
<td>Sum score UPDRS-ME (mean (SD, range))</td>
<td>37.7 (19.3; 7-82)</td>
</tr>
<tr>
<td>Dopaminergic therapy</td>
<td>45</td>
</tr>
<tr>
<td>Dyskinesias during videotape</td>
<td>1</td>
</tr>
</tbody>
</table>

YRS = years; SD = standard deviation; MHY = Modified Hoehn Yahr; UPDRS-ME = Unified Parkinson’s Disease Rating Scale motor examination.

Inter-rater variability of the UPDRS-ME

ICC’s and $\kappa_w$ are displayed in Table 2. For the nurses inter-rater reliability ranged from good to very good for the majority of individual UPDRS-items ($\kappa_w$ 0.65-0.93). Facial expression (0.43), head tremor (0.53-0.31), action tremor of the right arm (0.53-0.45), and action tremor of the left arm (0.55) had fair to moderate agreement. The inter-rater reliability for the sum score of the UPDRS-ME was very good (0.90-0.91). For the residents inter-rater reliability ranged from good to very good for the majority of individual UPDRS-items (0.60-0.94). Only three items had a fair to moderate agreement; facial expression (0.40), tremor right leg (0.41), and action tremor right arm (0.56). The kappa for head tremor for resident 1 could not be calculated. The inter-rater reliability for the sum score of the UPDRS–ME was very good (0.86-0.90). For the movement disorder specialist inter-rater reliability ranged from good to very good for the majority of individual UPDRS-items (0.66-0.91). Action tremor right arm (0.52) had a fair to moderate agreement. The inter-rater reliability for the sum score of the UPDRS–ME was very good (0.91).
Table 2. Inter-rater reliability of the UPDRS-ME: Intraclass correlation coefficients for the sum scores and weighted kappa values for the individual items. All raters compared to the senior MDS.

<table>
<thead>
<tr>
<th></th>
<th>Nurse 1</th>
<th>Nurse 2</th>
<th>Resident 1</th>
<th>Resident 2</th>
<th>MDS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sum score</td>
<td>.91</td>
<td>.90</td>
<td>.90</td>
<td>.86</td>
<td>.91</td>
</tr>
<tr>
<td>Speech</td>
<td>.83</td>
<td>.69</td>
<td>.78</td>
<td>.63</td>
<td>.72</td>
</tr>
<tr>
<td>Facial expression</td>
<td>.43</td>
<td>.66</td>
<td>.40</td>
<td>.66</td>
<td>.66</td>
</tr>
<tr>
<td>Tremor at rest</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>head</td>
<td>.53</td>
<td>.31</td>
<td>X</td>
<td>.70</td>
<td>.70</td>
</tr>
<tr>
<td>arm right</td>
<td>.81</td>
<td>.88</td>
<td>.64</td>
<td>.80</td>
<td>.80</td>
</tr>
<tr>
<td>arm left</td>
<td>.75</td>
<td>.78</td>
<td>.68</td>
<td>.75</td>
<td>.81</td>
</tr>
<tr>
<td>leg right</td>
<td>.90</td>
<td>.64</td>
<td>.61</td>
<td>.41</td>
<td>.66</td>
</tr>
<tr>
<td>leg left</td>
<td>.74</td>
<td>.68</td>
<td>.63</td>
<td>.65</td>
<td>.77</td>
</tr>
<tr>
<td>Action tremor</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>arm right</td>
<td>.53</td>
<td>.45</td>
<td>.56</td>
<td>.60</td>
<td>.52</td>
</tr>
<tr>
<td>arm left</td>
<td>.66</td>
<td>.55</td>
<td>.64</td>
<td>.64</td>
<td>.75</td>
</tr>
<tr>
<td>Finger tap</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>right</td>
<td>.72</td>
<td>.77</td>
<td>.86</td>
<td>.78</td>
<td>.76</td>
</tr>
<tr>
<td>left</td>
<td>.72</td>
<td>.83</td>
<td>.78</td>
<td>.76</td>
<td>.71</td>
</tr>
<tr>
<td>Hand movement</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>right</td>
<td>.69</td>
<td>.77</td>
<td>.71</td>
<td>.66</td>
<td>.69</td>
</tr>
<tr>
<td>left</td>
<td>.74</td>
<td>.85</td>
<td>.82</td>
<td>.72</td>
<td>.78</td>
</tr>
<tr>
<td>Pronation/supination</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>right</td>
<td>.70</td>
<td>.71</td>
<td>.73</td>
<td>.63</td>
<td>.74</td>
</tr>
<tr>
<td>left</td>
<td>.83</td>
<td>.80</td>
<td>.75</td>
<td>.71</td>
<td>.75</td>
</tr>
<tr>
<td>Leg agility</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>right</td>
<td>.69</td>
<td>.72</td>
<td>.85</td>
<td>.79</td>
<td>.87</td>
</tr>
<tr>
<td>left</td>
<td>.74</td>
<td>.71</td>
<td>.83</td>
<td>.75</td>
<td>.79</td>
</tr>
<tr>
<td>Arising from chair</td>
<td>.92</td>
<td>.91</td>
<td>.88</td>
<td>.94</td>
<td>.91</td>
</tr>
<tr>
<td>Posture</td>
<td>.65</td>
<td>.84</td>
<td>.67</td>
<td>.84</td>
<td>.83</td>
</tr>
<tr>
<td>Gait</td>
<td>.90</td>
<td>.84</td>
<td>.87</td>
<td>.87</td>
<td>.90</td>
</tr>
<tr>
<td>Postural stability</td>
<td>.93</td>
<td>.87</td>
<td>.85</td>
<td>.89</td>
<td>.91</td>
</tr>
<tr>
<td>Body bradykinesia</td>
<td>.72</td>
<td>.67</td>
<td>.69</td>
<td>.75</td>
<td>.77</td>
</tr>
</tbody>
</table>

UPDRS-ME = Unified Parkinson’s Disease Rating Scale motor examination; MDS = movement disorders specialist. X = weighted kappa could not be calculated because of missing values.

Using more detailed analysis, the Bland Altman plots (Figure 1) showed that both nurses, residents, and the MDS consistently assigned higher scores than the senior movement disorder specialist. Mean differences ranged between 1.7 and 5.4 (Table 3) (all differences p<0.05) with rather wide 95% agreement intervals. In addition a pattern of growing
disagreement with increasing values was observed between nurse 1 and senior MDS specialist. With respect to resident 1 difference tended to be in opposite directions for low and high UPDRS sum scores: for low UPDRS sum scores resident 1 assigned higher scores than the MDS senior, and for higher UPDRS sum scores lower scores.

Figure 1. Plots of differences between two raters against their mean for the total UPDRS-ME score. The mean value (dotted line), regression line of the differences (drawn line) and 95% limits of agreement (dashed line) are drawn. All mean differences were statistically significant different from 0 ($P<0.05$). This means that both nurses, residents, and the MDS consistently assigned higher scores than the senior MDS. Only the regression line of resident 1 (plot C) was statistically significant (slope B is $-0.22$, $P=0.001$). With respect to resident 1 difference tended to be in opposite directions for low and high UPDRS sum scores.
Intra-rater variability of the UPDRS-ME

Intra-rater reliability was assessed for nurse 1, resident 2, as well as the senior MDS. Reliability coefficients, ICC’s and $\kappa_w$ are displayed in Table 4. For the nurse practitioner intra-rater reliability ranged from good to very good for all individual UPDRS-ME items (0.60-0.94). The intra-rater reliability for the sum score of the UPDRS–ME was very good (0.91). For the resident intra-rater reliability ranged from good to very good for the majority of the individual UPDRS-ME items (0.66-0.97). Tremor right leg had a fair to moderate agreement (0.46). The intra-rater reliability for the sum score of the UPDRS–ME was very good (0.97). For the senior movement disorder specialist intra-rater reliability ranged from good to very good for the majority of individual UPDRS-items (0.74-0.94). Facial expression (0.58) and action tremor of the right and left arm (0.54-0.51) had a moderate agreement. The intra-rater reliability for the sum score of the UPDRS–ME was very good (0.95).

Additionally, 95% repeatability limits (or intra-rater agreement limits) were calculated (Table 3). The 95% repeatability limits of nurse 1 were comparable to the corresponding inter-rater 95% limits of agreement. (-6.8–11.0 vs. –8.5–11.9 respectively) For resident 2 the 95% repeatability interval was smaller than the corresponding inter-rater 95% agreement interval (-7.7–5.9 vs. –5.1-15.8) due to a considerably lower upper limit. The 95% repeatability limits of the senior MDS, –8.3 and 7.1, were more or less similar to the repeatability limits of resident 2.

### Table 3. Inter-rater variability for the total UPDRS-ME scores assigned by the nurses, residents, MDS and the senior MDS as well as 95% intra-rater agreement limits (or 95% repeatability limits).

<table>
<thead>
<tr>
<th></th>
<th>Mean difference</th>
<th>SD difference</th>
<th>95% CI of mean</th>
<th>P</th>
<th>95% inter-rater agreement limits</th>
<th>95% intra-rater agreement limits (or repeatability limits)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nurse 1</td>
<td>1.7</td>
<td>5.2</td>
<td>0.1 – 3.3</td>
<td>0.041</td>
<td>-8.5 – 11.9</td>
<td>-6.8 – 11.0</td>
</tr>
<tr>
<td>Nurse 2</td>
<td>4.1</td>
<td>4.3</td>
<td>2.8 – 5.5</td>
<td>&lt;0.001</td>
<td>-4.3 – 12.5</td>
<td>n.a.</td>
</tr>
<tr>
<td>Resident 1</td>
<td>1.9</td>
<td>4.9</td>
<td>0.3 – 3.5</td>
<td>0.019</td>
<td>(-0.7- 0.2<em>total UPDRS ME) – (15.7-0.2</em>total UPDRS ME)</td>
<td>n.a.</td>
</tr>
<tr>
<td>Resident 2</td>
<td>5.4</td>
<td>5.3</td>
<td>3.7 – 7.0</td>
<td>&lt;0.001</td>
<td>-5.1 – 15.8</td>
<td>-7.7 – 5.9</td>
</tr>
<tr>
<td>MDS</td>
<td>4.2</td>
<td>4.2</td>
<td>2.9 – 5.5</td>
<td>&lt;0.001</td>
<td>-4.1 – 12.5</td>
<td>n.a.</td>
</tr>
<tr>
<td>Senior MDS</td>
<td>reference</td>
<td></td>
<td></td>
<td></td>
<td>-8.3 – 7.1</td>
<td></td>
</tr>
</tbody>
</table>

*a Regression based limits of agreement since difference changes with magnitude of average scores. Total UPDRS-ME is the average score of both raters. In clinical practice, with only one rater, this is the score provided by that specific rater. b n.a.: not assessed.

Intra-rater variability of the UPDRS-ME

Intra-rater reliability was assessed for nurse 1, resident 2, as well as the senior MDS. Reliability coefficients, ICC’s and $\kappa_w$ are displayed in Table 4. For the nurse practitioner intra-rater reliability ranged from good to very good for all individual UPDRS-ME items (0.60-0.94). The intra-rater reliability for the sum score of the UPDRS–ME was very good (0.91). For the resident intra-rater reliability ranged from good to very good for the majority of the individual UPDRS-ME items (0.66-0.97). Tremor right leg had a fair to moderate agreement (0.46). The intra-rater reliability for the sum score of the UPDRS–ME was very good (0.97). For the senior movement disorder specialist intra-rater reliability ranged from good to very good for the majority of individual UPDRS-items (0.74-0.94). Facial expression (0.58) and action tremor of the right and left arm (0.54-0.51) had a moderate agreement. The intra-rater reliability for the sum score of the UPDRS–ME was very good (0.95). Additionally, 95% repeatability limits (or intra-rater agreement limits) were calculated (Table 3). The 95% repeatability limits of nurse 1 were comparable to the corresponding inter-rater 95% limits of agreement. (-6.8–11.0 vs. –8.5–11.9 respectively) For resident 2 the 95% repeatability interval was smaller than the corresponding inter-rater 95% agreement interval (-7.7–5.9 vs. –5.1-15.8) due to a considerably lower upper limit. The 95% repeatability limits of the senior MDS, –8.3 and 7.1, were more or less similar to the repeatability limits of resident 2.
Discussion

This is a comparative evaluation of the inter-rater and intra-rater variability of the UPDRS-ME of most of the health care professionals involved in the care of PD patients, i.e. nurse practitioners, residents in neurology and movement disorder specialists. Presence
of systematic difference in assigned UPDRS-ME scores between these different raters compared to a senior MDS was observed for the whole range of scores when plotting inter-rater differences against their mean. The limits of agreement show that if patients are assessed by two different observers, differences in the assigned UPDRS-ME scores can be as large as 16 points (referring to the largest of the lower and upper limits of agreement). This means that if patients are assessed on two different occasions by two different observers, real changes in the total UPDRS-ME scores of less than 16 points cannot be distinguished from measurement error. This considerable inter-rater disagreement can partly be attributed to the rather large intra-rater variation. This lack of agreement was by no means obvious in the kappa and ICC values for the UPDRS-ME scores, which suggested good inter-rater and intra-rater reliability.

Evaluation of rater variability only by kappa and ICC’s is insufficient. Measures of agreement as derived by Bland Altman plots provide insight into the ability of two or more observers to achieve the same value, while kappa statistics and ICC’s indicate the ability of two or more observers to differentiate between subjects in a group.(23) The main weakness of kappa is that it fails to detect systematic difference between raters.(24) The ICC is defined as the ratio of the variance between patients over the total variance. A higher value indicates that less variance is due to different sources of error, such as differences between observers. However, the actual amount of disagreement between observers is not indicated by the ICC.

Previous reports on rater variability of the UPDRS-ME in PD patients only considered kappa and ICC coefficients.(3,7,8,10,11) In these studies intra-rater and inter-rater reliability for the individual items ranged from poor to good. This inconsistency concerned the core features bradykinesia, tremor and rigidity (6) and may be explained by small sample sizes (7,10), use of different populations, e.g. normal versus older population (8), as well as the use of a modified versions of the UPDRS-ME.(8,10) The present patient sample, in contrast, enclosed PD patients with a wide range of age distribution, varying disease duration, and all degrees of disease severity in terms of UPDRS-ME sum scores and disability in terms of Hoehn and Yahr stages are present. The patient sample did not enclose PD patients with dyskinesias. Accordingly, our results are applicable to the clinical setting in PD patients without dyskinesias.

Systematic difference between the raters and the senior MDS occurred despite a prior training session to ensure similarity of rating methods. This training consisted of an explanation of the UPDRS-ME by an expert, scoring and discussion of 10 videotaped UPDRS-ME administrations as well as the UPDRS-ME Teaching Tape. This videotape is widely accepted and recommended as a training device to standardise the practical application of the UPDRS-ME and guarantee uniform interpretation.(6,9) Recently Goetz and Stebbins reported on the utility of the UPDRS Teaching Tape for successful training of
The UPDRS-ME administration. They examined among 226 raters the proportion that successfully rated the UPDRS-ME within 95% confidence interval range established by three experts. It appeared that only 55% of raters scored within the acceptable range on their first rating. Passing rates did not differ by professional status (half of the raters were university professors of neurology with expertise in PD). This is in accordance with our finding of systematic difference between the senior MDS and both nurses and residents as well as a MDS. This implies that the ability of nurses and residents in neurology to administer the UPDRS does not differ from the ability of a MDS.

In the same report the most difficult cases of the UPDRS-ME teaching tape appeared the cases with minimal and severe Parkinsonism, for whom respectively 98% of the failures were due to overrating and 83% of the failures were due to underrating. In the present analysis rater agreement in the subgroup of patients with minimal PD (cohort) was comparable to the rater agreement in the subgroup with advanced PD (trial) (results not shown).

A search for possible problematic items that could account for the systematic difference in total UPDRS-ME scores between the senior MDS and the other raters did not yield an answer. The systematic differences appeared to be present for all items of the UPDRS-ME. The systematic difference cannot be attributed to different scores in severity subgroups or one of the different items of the UPDRS-ME. We hypothesize that the fact that the senior MDS is predominantly engaged in the care of the most serious, complicated type of PD patients may have contributed to his systematically lower scoring of the present patients. The presence of dyskinesia may interfere with the assessment of the UPDRS-ME. Because dyskinesias mainly occur during the on-state, the 20 mild to moderate PD patients in our sample, from the ongoing cohort study, were at risk for developing dyskinesias. As only one of them had mild dyskinesias, this will not have influenced our results.

In conclusion, our study shows considerable observer difference for the whole range of UPDRS-ME scores between a senior MDS and nurse practitioners, residents in neurology, and a MDS. This could in part be attributed to within observer disagreement. This implies that investigators should quantify the agreement of their assessments before starting longitudinal studies of disease progression or clinical trials. Ideally, the same observer should be responsible for the assessment of each patient, since intra-rater agreement is in general better than inter-rater agreement. This applies even more for individual patient management, since the reliability of an instrument for individual patient management should be higher than for use in group comparisons in clinical trials. Finally, evaluation of rater agreement should not be limited to $\kappa$ and ICC statistics but always include the assessment of the extent of bias between different raters.
Reference List


