Unified Parkinson's Disease Rating Scale-Motor Exam: Inter-rater reliability of advanced practice nurse and neurologist assessments

Janice L. Palmer, PhD, RN[Assistant Professor],
Saint Louis University School of Nursing, 3525 Caroline Mall, Room 326, St. Louis, MO, 63104, USA

Mary A. Coats, MSN, RN[Research Associate Professor],
Alzheimer's Disease Research Center at Washington University, 4488 Forest Park, Suite 101, St. Louis, MO, 63108, USA

Catherine M. Roe, PhD[Instructor of Neurology],
Washington University School of Medicine, 660 S. Euclid Avenue, Box 8111, St. Louis, MO, 63110, USA

Shelly M. Hanko, PhD, FNP-BC[Assistant Professor],
University of Missouri-St. Louis College of Nursing, One University Boulevard, 225 Seton Hall, St. Louis, MO 63121, USA

Chengjie Xiong, PhD[Associate Professor of Biostatistics], and
Division of Biostatistics, Washington University, 660 S. Euclid Avenue, St. Louis, MO, 63110, USA

John C. Morris, MD[Professor][Director]
Alzheimer's Disease Research Center at Washington University, 4488 Forest Park, Suite 101, St. Louis, MO, 63108, USA

Abstract

Aim—This paper is a report of a study to establish the inter-rater reliability of advanced practice nurse and neurologist neurological assessments which included ratings with the Unified Parkinson's Disease Rating Scale-Motor Exam.

Background—Around the world, advanced practice nurses are performing tasks once completed by only physicians. To promote consumer and provider confidence, it is important to establish that nurse and physician ratings using assessment tools are similar. In addition in research settings,

when different raters are used, establishment of inter-rater reliability for study assessments is needed.

Method—Advanced practice nurses and neurologists independently recorded findings on neurological examinations of 46 participants in a study conducted between August 2007 and January 2008. An intraclass correlation coefficient was calculated to estimate overall agreement between the nurse and neurologist ratings. Agreement for individual items measured on a dichotomous scale was assessed by calculating Cohen's kappa.

Results—There was substantial agreement between advanced practice nurses and neurologists on the mean Unified Parkinson's Disease Rating Scale-Motor Exam ratings (intraclass correlation coefficient = 0.65) and the U.S. National Alzheimer's Coordinating Center Uniform Data Set neurological examination ratings of unremarkable findings (kappa = 0.74) and of gait disorder (kappa = 0.73). Moderate agreement (kappa = 0.53) was reached for the rating of whether all Unified Parkinson's Disease Rating Scale-Motor Exam items were normal.

Conclusion—These findings are consistent with studies of the inter-rater agreement of the Unified Parkinson's Disease Rating Scale-Motor Exam and support the conduct of neurological assessments by advanced practice nurses.

Keywords

INTRODUCTION

Advanced practice nurses around the world are performing tasks once thought able to be completed only by physicians. When using assessment tools, findings of nurse practitioners or clinical nurse specialists must be reliable and similar to those of physicians. This report is of a study comparing advanced practice nurse and neurologist ratings on neurological examinations and ratings using the Unified Parkinson's Disease Rating Scale-Motor Exam (UPDRS-ME). This scale is used in clinical practice as a measure of Parkinson disease progression. In addition, data from the UPDRS-ME and Uniform Data Set (to be described later) are included in the U.S. National Alzheimer's Coordinating Center database for Alzheimer disease research. Ideally, ratings within and between research centers contributing to the national database should be reliable. One of the contributing research centers was the setting for this study.

BACKGROUND

Agreement between advanced practice nurse ratings and physician ratings has been reported for other assessment instruments. For example, a nurse practitioner's and cardiologist's interpretation of 100 exercise stress test (EST) findings was found to be moderate to high (Maier *et al.* 2008). In none of the 100 cases did the nurse practitioner's conclusion about whether the EST was normal (positive or negative finding) differ from the cardiologist's determination. The researchers concluded that the study provided support for the experienced nurse practitioner's role to include interpretation of ESTs (Maier *et al.* 2008). A recent study of inter-rater reliability between a physician and a nurse practitioner on the Wells score in the assessment of deep vein thrombosis in an emergency department showed that for 81 of the 100 cases, the nurse practitioner assessments resulted in the same Wells score as the physician (kappa=.74; Dewar & Corretge 2008). Good inter-rater reliability was shown, with the Wells score being described as a reliable tool for assessment regardless of assessor (Dewar & Corretge 2008).

Intra-rater and inter-rater agreement for nurses and physicians have also been reported for the UPDRS (Fahn & Elton 1987), which was one of the measures of interest for this study. Bennett et al. (1997) reported consistency (ranging from 0.70 to 0.95 on the individual domains and 0.90 to 0.95 for the total score) in nurse ratings of the UPDRS, repeated with the same patients (N=75) approximately 18 days later. In the same study, nurses had good to excellent agreement with a neurologist's ratings (Bennett 1997). However, Post et al. (2005) reported considerable disagreement when the ratings of senior movement disorder specialists were compared to those of a less experienced movement disorder specialist, two nurse practitioners, and two neurology physicians on 50 videotaped "off" state recordings of UPDRS-ME assessments. The intraclass correlations of the sum score ranged from 0.86 to 0.91. Inter-rater reliabilities (kappas) of the individual items ranged from 0.31 to 0.92. Intra-rater reliability for the sum score ranged from 0.91 to 0.97 and for individual items ranged from 0.46 to 0.94. The researchers concluded that inter-rater reliability should be determined for each study site before conducting longitudinal studies, and when possible the same rater should follow research participants in longitudinal studies because intra-rater reliability was better than inter-rater reliability (Post et al. 2005). The intra-rater reliability findings were similar to those reported for another study with movement disorder specialists (Siderowf *et al.*. 2002). A comparison of movement disorder specialists' ratings with their own repeat evaluation and rating of 404 clinical trial participants approximately two weeks later showed that the UPDRS total score reached an agreement level of 0.92, and the subscale scores ranged from 0.69 to 0.90 agreement (Siderowf *et al.* 2002). However, agreement on individual items such as intellectual impairment and rest tremor was as low as 0.49 (Siderowf *et al.* 2002).

In 2005, to promote the comparison and use of data between study sites, the U.S. National Alzheimer's Coordinating Center (NACC) implemented collection of a Uniform Data Set (UDS) for all Alzheimer's Disease Centers (ADCs) funded by the National Institute on Aging (NIA). The Unified Parkinson's Disease Rating Scale-Motor Exam (UPDRS-ME) was an assessment instrument included in the UDS battery (Morris *et al.* 2006). Based on concerns by the researchers and site-specific external grant reviewers at one of the ADCs about the comparability of advanced practice nurse and neurologist findings when completing the UDS battery assessments, a study was conducted between August 2007 and January 2008 to determine inter-rater reliability in performing and documenting findings on the UPDRS-ME and the neurological examination components.

THE STUDY

Aim

The aim of the study was to establish the inter-rater reliability of advanced practice nurse and neurologist neurological assessments which included ratings with the Unified Parkinson's Disease Rating Scale-Motor Exam.

Design

An instrument validation study was conducted. Intra-rater reliability is a measure of the consistency of one examiner on repeat assessments, and inter-rater reliability is a measure of the consistency that two or more individuals will have the same findings with the same assessment. The strength of agreement between different raters, or between the same rater at different times, can be approximated numerically using statistical tests such as Cohen's kappa (Cohen, 1960) and intraclass correlation coefficient [ICC] (Strout & Fleiss 1979). The strength of agreement yielded by these tests can range from 0.0-1.0. The closer the value is to 1.0, the stronger the agreement. We used the Landis and Koch (1977) "arbitrary" "benchmarks" for strength of agreement for interpreting kappa values and ICCs with values

of 0.21 to 0.40 indicating fair, 0.41 to 0.60 moderate, 0.61 to 0.80 substantial, and greater than 0.81 suggesting "almost perfect" correlation (p. 165). For this study, an ICC was calculated to estimate the overall agreement of advanced practice nurse and neurologist findings on the UPDRS-ME and a Cohen's kappa was calculated to estimate agreement on the ratings of each individual item.

Participants

One of the U.S. NIA-funded Alzheimer's Disease Research Centers which contributes to the NACC data repository was the setting for this study. Over 500 research participants are evaluated annually at this center by physicians and advanced practice nurses in a longitudinal study of Alzheimer's disease and healthy cognitive aging. The research participants are healthy volunteers who have mild cognitive loss and possibly a few mild symptoms of a movement disorder consistent with those expected with a dementing disorder such as Alzheimer disease and healthy volunteers with no known neurological disorders. To examine inter-rater reliability on the UPDRS-ME and neurological examination portions of the battery, a head-to-head comparison of nurse clinician ratings with neurologist ratings was conducted for 46 longitudinal study participants. In total, two nurse practitioners and two clinical nurse specialists, all with a minimum of seven years' experience in dementia assessment, and four neurologists participated as raters in this study. Although not ideal, assignment of rater pairs was not completely random but was necessarily adjusted according to the scheduling needs of the Center.

Training

Prior to the study, the nurses and neurologists watched and scored a video teaching tape developed by the Movement Disorder Society (2003) for training on the UPDRS-ME. This training was augmented by detailed discussion by the participating clinicians of the scoring of some items and repeated viewings of all or portions of the tape by the clinicians. Discussion led to consensus on ratings.

Training issues requiring repetition of the video sessions were similar to those previously reported (Goetz & Stebbins 2004). Goetz & Stebbins reported that for training on the UPDRS-ME for a clinical trial, three Parkinson disease experts rated four patients shown on a videotape. Their ratings were used to establish passing rates for clinical trial certification. Only 55% of the 226 raters passed their first rating of the four patients (Goetz & Stebbins 2004). All raters passed on their third viewing and rating attempt. There was no statistically significant difference in the passing rates of neurologists compared to other physicians and study coordinators (Goetz & Stebbins 2004).

Data Collection

Pairs of clinicians (an advanced practice nurse and a neurologist) independently and without discussion indicated whether each UDS neurological summary finding was present (yes, no) or unknown (See Table 1). They also assigned scores for each item of the UPDRS-ME; possible scores ranged from 0 to 4, with higher scores indicating greater abnormality.

Ethical Considerations

Institutional review board approval was received from the study institution and informed consent was obtained from the nurse and physician participants prior to the study. The primary risk included in the consent form and reviewed with the nurse and physician clinicians was possible anxiety about a comparison of their assessment findings to those of other clinicians. Because the advanced practice nurses were repeating the neurological examinations of regularly-scheduled neurologist assessments, another risk factor for the

nurses was an increase in work-load. Duplicate examination of the longitudinal research participants was covered in their existing consent forms.

Data Analysis

A mean UPDRS-ME rating of the 27 items was calculated for each participant/rater combination (i.e., each participant had two UPDRS-ME ratings, one by the nurse and one by the neurologist). Items scored as "untestable" were treated as missing, and the mean rating was calculated without missing items. An intraclass correlation coefficient (ICC; Shrout & Fleiss 1979) was calculated to estimate overall agreement between nurse and neurologist raters. In this analysis, ICC (1.1) was used, which assumes that each participant is rated by multiple raters and that all participants have the same number of raters (Shrout & Fleiss 1979). Agreement between the nurses and neurologists for individual items measured on a dichotomous scale: the overall "Normal/Abnormal" rating on the UPDRS-ME; the overall summary UDS neurological exam finding, whether or not focal deficits were present, and whether or not a gait disorder was present were assessed by calculating Cohen's kappa (Cohen 1960). In the computation of kappa statistics, nurses were treated as one rater and physicians were treated as the other rater.

RESULTS

Using the Landis and Koch (1977) guidelines, agreement between the nurses and physicians was *substantial* for the mean UPDRS-ME ratings (ICC=0.65; 95% CI=0.79-0.45), UDS neurological examination ratings of unremarkable findings (kappa=0.74), and ratings of gait disorder (kappa=0.73); *moderate* for whether all UPDRS-ME items were normal (kappa=0.53); and *fair* for UDS neurological examination ratings of focal deficits (kappa=0.23; See Table 1). The *fair* correlation on focal deficits was most probably due to the small number of participants with focal findings. Kappa could not be appropriately computed for the question, "Are there eye movement abnormalities present indicative of a central nervous system disorder?", because the number of participants was very unbalanced between cells in the cross-tabulation table used to compute the kappa statistic. For that question, the nurse and physician agreed for 45 of 46 participants that the person did not have eye movement abnormalities. For the remaining participant, the physician said "no" eye movement abnormalities were present and the nurse conservatively answered "unknown".

DISCUSSION

Study limitations

This study had several limitations. It was halted prematurely related to nurse staffing needs of clinical trials and other pressing projects of the research center. Furthermore, because of the nurses and physicians' schedules, assignment of pairings was not random. We believe that these limitations reflect real-world issues in conducting reliability studies of assessment tools, as research offices and other nurse practice settings are often busy with competing demands for nurse time. Having only 46 paired ratings that were not randomly assigned limits the statistical power to detect differences in ratings. Also, the sample being studied was comprised of volunteers who had no known neurological illness (control group) or participants who had mild cognitive loss and possibly a few mild symptoms of a movement disorder; therefore the spectrum of possible findings was limited by the inclusion and exclusion criteria of the larger research project. In addition, the limited number of raters— four advanced practice nurses and four neurologists—although consistent with the make-up and size of a real world clinical practice or research office, also limits the generalizability of

our findings. We used kappa and ICC to quantify the agreement between advanced practice nurses and neurologists, but other methods are also available.

Typical of inter-rater reliability studies, training and discussion of scoring guidelines on the structured assessment tools occurred prior to study initiation. This structure and additional training is likely to have had an impact on the generalizability of the findings. Also limiting generalizability is the fact that the study site differs from other clinical and research settings with regard to the scope of practice, education, and certification of the advanced practice nurses.

However, we believe that the substantial level of agreement for the mean UPDRS-ME rating, UDS neurological examination ratings of unremarkable findings and gait disorder provides initial support for nurse practitioners and clinical nurse specialists in performing the NACC UDS neurological examination and the UPDRS-ME. The reliability findings for the UPDRS-ME are consistent with those of other inter-rater reliability studies of the UPDRS-ME, suggesting that nurse practitioners and clinical nurse specialists can assess and score the UPDRS-ME similarly to physicians (Bennet *et al.* 1997, Fahn & Elton 1987, Post *et al.* 2005).

Reliability of the Unified Parkinson's Disease Rating Scale-Motor Exam

This study provides additional evidence that the reliability of the UPDRS-ME instrument is less than ideal, in that the kappa value for the rating of whether all items were normal demonstrated only moderate agreement. Concerns related to inter-rater reliability and other limitations of the UPDRS led a 2003 Movement Disorder Society Task Force on Rating Scales for Parkinson disease to recommend modifications to the UPDRS (Movement Disorders Society 2003). A revised UPDRS, the MDS-UPDRS is available at the Movement Disorders Society website (www.movementdisorders.org; Goetz *et al.* 2007). As of this writing (January 22, 2010), the U.S. National Alzheimer's Coordinating Centers continue to use the UPDRS-ME version reported on in this paper for the Uniform Data Set (Personal communication, M.A. Coats, January 22, 2010).

CONCLUSION

The moderate level of agreement on whether all the UPDRS-ME items were normal and the substantial level of agreement on the mean UPDRS-ME ratings and NACC UDS summary neurological findings support the continuation of assessments by advanced practice nurses on these measures at this research setting.

Our results support those of other studies that advanced practice nurses can perform to the gold standard, given adequate training in the scope of practice under investigation. Policies on training of all individuals, regardless of profession or certification level, on the administration of an assessment tool to achieve rater agreement and limit bias should be implemented prior to use of the assessment tool in research or with the introduction of a new practitioner into a practice setting. Future studies are needed in other settings with this and other tools to establish advanced practice nurse and physician inter-rater reliability. Such research will help further define the advanced practice nurse role in practice.

SUMMARY STATEMENT

What is already known about this topic

- Agreement between advanced practice nurse and physician ratings has been reported for several measures, including interpretation of exercise stress tests and assignment of Wells scores for deep vein thrombosis.

Previous studies of inter-rater reliability for the Unified Parkinson's Disease Rating Scale-Motor Exam have had mixed results.

What this paper adds

- There was substantial agreement between advanced practice nurses and neurologists on mean Unified Parkinson's Disease Rating Scale-Motor Exam ratings and the Uniform Data Set summary neurological examination ratings of unremarkable findings and ratings of gait disorder.

- Moderate agreement was reached for the Unified Parkinson's Disease Rating Scale-Motor Exam rating of whether all items were normal.

Implications for practice and/or policy

- The substantial level of agreement for the mean Unified Parkinson's Disease Rating Scale-Motor Exam rating and Uniform Data Set summary neurological examination ratings of unremarkable findings and gait disorder provide support for the ability of nurse practitioners and clinical nurse specialists in performing neurological examinations.

- The reliability of the Unified Parkinson's Disease Rating Scale-Motor Exam is less than ideal.

- Assessments using the Unified Parkinson's Disease Rating Scale-Motor Exam can be reliably conducted by advanced practice nurses.

Acknowledgments

REFERENCES

Bennett DA, Shannon KM, Beckett LA, Goetz CG, Wilson RS. Metric properties of nurses' ratings of parkinson signs with a modified Unified Parkinson's Disease Rating Scale. Neurology. 1997; 49(6): 1580–1587. [PubMed: 9409350]

Cohen J. A coefficient of agreement for nominal scales. Educational and Psychological Measurement. 1960; 20:37–46.

Dewar C, Corretge M. Inter-rater reliability of the Wells score as part of the assessment of DVT in the emergency department: Agreement between consultant and nurse practitioner. Emergency Medical Journal. 2008; 25(7):407–410.

Fahn, S.; Elton, RL. UPDRS Development Committee. The Unified Parkinson's Disease Rating Scale. In: Fahn, S.; Marsden, CD.; Calne, DB.; Goldstein, M., editors. Recent Developments in Parkinson's Disease. 2nd edn. Macmillan Healthcare Information; Florham Park, NJ: 1987. p. 153-163.p. 293-304.

Goetz CG, Stebbins G. Assuring interrater reliability for the UPDRS motor section: Utility of the UPDRS teaching tape. Movement Disorder. 2004; 19(12):1453–1456.

Goetz CG, Fahn S, Martinez-Martin P, Poewe W, Sampais C, Stebbins GT, Stern MB, Tilley BC, Dodel R, Dubois B, Holloway R, Jankovic J, Kulisevsky J, Lang AE, Lees A, Leurgens S, LeWitt PA, Nyenhuis D, Olanow CW, Rascol O, Schrag A, Teresi JA, van Hilten, LaPelle N. Sponsored revision of the Unified Parkinson's Disease Rating Scale (MDS-UPDRS): Process, format, and clinimetric testing plan. Movement Disorder. 2007; 22(1):41–47.

Landis JR, Koch GG. The measure of observer agreement for categorical data. Biometrics. 1977; 3:159–174. [PubMed: 843571]

Maier E, Jensen L, Sonnenberg B, Archer S. Interpretation of exercise stress test recordings: Concordance between nurse practitioner and cardiologist. Heart & Lung. 2008; 37(2):144–152. [PubMed: 18371507]

Morris JC, Weintraub S, Chui H, Cummings J, DeCarli C, Ferris S, et al. The Uniform Data Set (UDS): Clinical and cognitive variables and descriptive data from Alzheimer Disease Centers. Alzheimer Disease Associated Disorder. 2006; 20(4):210–216.

Movement Disorder Society Task Force on Rating Scales for Parkinson's Disease. The Unified Parkinson's Disease Rating Scale (UPDRS): Status and recommendations. Movement Disorder. 2003; 18(7):738–750.

Post B, Merkus M, de Bie RMA, de Haan RJ, Speelman JD. Unified Parkinson's Disease Rating Scale Motor Examination: Are ratings of nurses, residents in neurology, and movement disorders specialists interchangeable? Movement Disorder. 2005; 20(12):1577–1584.

Shrout PE, Fleiss JL. Intraclass correlations: Uses in assessing rater reliability. Psychological Bulletin. 1979; 86:420–428. [PubMed: 18839484]

Siderowf A, McDermott M, Kieburtz K, Blindauer K, Plumb J, Shoulson I, et al. Test-retest reliability of the Unified Parkinson's Disease Rating Scale in patients with early Parkinson's disease: Results from a multicenter clinical trial. Movement Disorder. 2002; 17(4):758–763.

TABLE 1

Agreement between Advanced Practice Nurses and Neurologists on the Unified Parkinson's Disease Rating Scale-Motor Exam and the U.S. National Alzheimer's Coordinating Center Uniform Data Set Neurological Exam

Measure	Kappa	95% CI Lower	Upper
UPDRS-ME, all items normal	.527	.289	.766
Uniform Data Set Neurological Exam			
1. Are all findings unremarkable (normal or abnormal for age)?	.744	.508	.981
2. Are focal deficits present indicative of central nervous system disorder?	.228	−.230	.687
3. Is gait disorder present indicative of central nervous system disorder?	.728	.373	1.000
4. Are there eye movement abnormalities present indicative of central nervous system disorder?	See Results section		