Published in final edited form as:

Parkinsonism Relat Disord. 2013 February ; 19(2): 232–237. doi:10.1016/j.parkreldis.2012.10.015.

Validating an objective video-based dyskinesia severity score in Parkinson's disease patients

Anusha Sathyanarayanan Rao¹, Benoit M. Dawant¹, Robert E. Bodenheimer¹, Rui Li¹, John Fang², Fenna Phibbs², Peter Hedera², and Thomas Davis^{2,*}

¹Department of Electrical Engineering and Computer Science, Vanderbilt University, Nashville, Tennessee, USA

²Department of Neurology, School of Medicine, Vanderbilt University, Nashville, Tennessee, USA

Abstract

Dyskinesia is a common side effect of prolonged dopaminergic therapy in Parkinson's disease patients. Assessing the severity of dyskinesia can help develop better pharmacological and surgical interventions. We have developed a semi-automatic video-based objective dyskinesia quantifying measure called the severity score (SVS) that was evaluated on 35 patient videos. We present a study to evaluate the utility of our severity score and compare its performance to clinical ratings of neurologists. In addition to the Unified Dyskinesia Rating Scale (UDysRS) score for each video, four neurologists provided three sets of time lapsed ratings and rankings of the 35 videos using a specifically developed protocol. The statistical analysis of our data using Kendall's tau-b and intra-class correlations shows that (a) ranking patient videos based on severity is suitable for studying the utility of the SVS, and (b) SVS exhibits moderate utility to quantify dyskinesia severity when compared to manual assessment of dyskinesia by neurologists using the UDysRS. These results support the effective use of SVS as an objective measure to quantify dyskinesia and the rationale for a ranking system that complements traditional rating scales.

Keywords

Parkinson's disease; rating scales; quantify dyskinesia; rater variability; severity ranking

INTRODUCTION

Levodopa therapy in Parkinson's disease (PD) patients results in drug-induced dyskinesia characterized by hyperkinetic involuntary movements that may often interfere with activities of daily living [1]. Despite current treatment measures, the disabling symptoms of dyskinesia continue to challenge the development of better pharmacological and surgical interventions. In this context, rating scales have been the most established and widely used

*Corresponding author: Tel: (615) 936-0060; Fax: (615) 936-1229, thomas.l.davis@Vanderbilt.Edu (Thomas L. Davis).

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

means of assessment of the severity of dyskinesia. The key attributes of dyskinesia evaluated include anatomical distribution, phenomenology, duration, intensity, disability, and patient perception [2]. Different scales base their severity ratings on different sets of attributes of dyskinesia. Some of the widely used ratings scales are the Abnormal Involuntary Movement Scale (AIMS)[3], the Lang Fahn activities of daily living scale [4], the Rush Dyskinesia Rating Scale[5], the Parkinson's disease Dyskinesia Scale[6], and the Clinical Dyskinesia Rating Scale (CDRS) [7]. The most recently developed scale is the Unified Dyskinesia Rating Scale (UDysRS), which may become the standardized dyskinesia rating scale equivalent to the UPDRS scale for PD symptoms [2]. The UDysRS is a combination of several rating scales in such a way that all attributes of dyskinesia are assessed using a single rating scale. The results of the clinimetric testing of this scale over a range of 70 patients indicated an inter-rater and intra-rater reliability with correlation coefficients ranging from 0.37 to 0.87 for various tasks. Further validation and responsiveness testing is underway [2].

Though rating scales are the conventional assessment tool, there are several disadvantages to their use [8]. First, they are subjective and often require intensive training to obtain acceptable intra- and inter-rater reliability. Second, scales such as the CDRS and UDysRS include patient questionnaires, which may not represent the severity of dyskinesia accurately. The rating scales often rely on a discrete five point scale. This lack of resolution leads to the possibility of misclassifying patients with symptoms that fall in between two rating intervals. These factors encourage the development of quantitative assessment techniques, which has been our primary research interest. Accelerometers and gyroscopes have achieved moderate success in quantifying the severity of dyskinesia in the recent past [9, 10, 11, and 12]. The disadvantages of these techniques are the use of expensive and dedicated devices that require complex software, and the inconvenience to the patients wearing the devices.

Our work is an example of a video-based, marker-less, model-free human motion tracking using the standardized clinical videos of PD patients that are often a part of the patient's clinical records. Using this technique, we have developed a severity score (SVS) to quantify the severity of dyskinesia exhibited by the patient [13]. As a continuous variable, SVS cannot be directly compared to the discrete rating scales. We therefore developed a rating based ranking protocol to validate the utility of SVS to quantify dyskinesia. Each patient in the study had the following parameters: neurologist ratings and rankings, UDysRS rankings and SVS ranking. Three studies were performed using these parameters to establish the utility of the SVS and the ranking protocol – (a) Assessment of intra- and inter-neurologist consistency using the ranking protocol; (b) Comparison of SVS with the UDysRS and the neurologists; and (c) Effect of ratings vs. rankings. Our results indicate that SVS correlates well with neurologists' and the UDysRS rankings. The validity of the SVS was studied by evaluating it using different video segments of some of the patient videos. This longitudinal study on the videos was performed to observe if 10s of video data was sufficient to quantify dyskinesia severity in patients.

METHODS

Our analysis used samples of 35 patient videos with varying dyskinesia severity obtained as part of the extensive clinimetric testing of the UdysRS [2]. The videos were captured in a controlled environment with plain backgrounds in a well-illuminated room with no occluding furniture. Details regarding the video protocol and the informed consent obtained have been previously published [2]. The patients were rated based on four tasks that are activities of daily living (ADL). Part IV UDysRS scores for each task was also available. Our task of interest was the communication task, where the patients were asked to read while seated on a chair. The communication task was the simplest task to track using our semi-automatic technique. Though speech disorders were primarily rated using this task, the patients also exhibited movement of the face, head, neck, hands and legs and it was observed that the patients with severely impaired speech also showed dyskinetic movements of these body parts. The average length of the communication task was one minute. A 10s excerpt from the middle of each 60s sequence was analyzed using our semi-automatic technique. Our previous work had determined that 10 seconds of video provided the optimal registration results without significant tracking delay. We avoided any starting and stopping movement effects by not analyzing the beginning or end of the video. The movements of patient's head, shoulders, chest, forearms, knees, feet, and the reading material were semi-automatically tracked using the Adaptive Bases Algorithm (ABA) that is an intensity based non-rigid image registration algorithm [14]. We analyzed the tracked anatomical points of interests by applying principal component analysis (PCA) on the cluster of points from every frame of the video sequence as described in our prior work [15]. A severity score was computed for each video sequence using the parameters obtained from the PCA analysis

$$SVS = TV * NSM/STDEV$$

TV: total variance of all eigenmodes, where the total variance is the sum of the magnitude of all the eigenvalues.

NSM: Number of significant modes of variation, which defines the number of modes of variations that capture 90% of the variations in the patient movements

STDEV: standard deviation of the percentage contribution of the eigenvalues to the total variance, which represents the rate of fall of eigenvalues. A gradual fall of eigenvalues indicates complex patient movements in various directions and a steep fall indicates simple movements in fewer directions.

The 10s sequences were ranked in the increasing order of SVS. These videos were also ranked based on the increasing order of the UDysRS part IV communication task scores obtained from the UDysRS study.

Four movement disorder neurologists, N1, N2, N3, and N4, ranked the 35 10s video sequences using a ranking protocol that included amplitude and speed of dyskinetic movement, anatomical distribution of dyskinesia and the extent of disability seen in the patient. Intelligibility of speech was not considered in the ranking. Each neurologist independently rated and ranked the 35 video segments using this protocol. Three sets of

ratings and rankings on the same dataset were obtained at monthly intervals to ensure they were not voluntarily repeated. Thus, each neurologist had three sets of ratings and rankings – ratings: Set 1R, Set 2R, and Set 3R and rankings: Set 1r, Set 2r, and Set 3r.

- a. Ratings: The videos were first rated on a scale of one to four with one - no dyskinesia, two - mild dyskinesia, three - moderate dyskinesia and four - severe dyskinesia.
- b. Rankings: The videos in each rating category, except the no dyskinesia category, were viewed simultaneously on a single screen and ranked according to increasing order of severity within that category.
- c. The first two and the final two videos in each rating category were then compared with the correspondingly ranked videos of the immediately next category to confirm if these rankings were still valid. Thus, the neurologist could view cross category videos to finalize their ranks.
- d. In case of rank changes, steps (b) and (c) were repeated until ranks were finalized and the corresponding rating categories in Step (a) were also modified to ensure coherence between ratings and ranking.

Each patient in the study had the following parameters: three sets of neurologist ratings and rankings, UDysRS ranking and SVS ranking.

The longitudinal study on the videos was performed by computing the SVS using the above technique on two more video segments on patients who had good longitudinal video data available. Five patients had good video data for 30 – 60s of the communication task and two 10s segments apart from the original segment were used for this study.

Data Analyses

Three statistical studies were conducted from the data obtained using the above methods.

- a. Validation of ranking protocol: Evaluation of intra- and inter-neurologist ranking consistency.
- b. SVS Utility: Comparison of SVS rankings with UDysRS rankings and the neurologists rankings
- c. Effect of ratings vs. rankings: Comparison of SVS with neurologists' ratings and rankings.

The original rankings obtained from neurologists were modified as follows to permit statistical analysis because the number of non-dyskinetic patients (neurologist rating of 1) was different in both the inter- and intra-neurologist ratings. Hence, the total number of patients ranked by each neurologist was not necessarily equal. To ensure statistical consistency in the analyses, for each set of rankings, two types of rank data sets were developed.

1. Type I: All 35 video sequences were part of this dataset. A tied rank was assigned to non-dyskinetic patients such that its value is the average of the ranks the patients would have received if there were given distinct ranks [16]. This process ensured

the maximum ranking in each ranking was 35, but the minimum rank would depend on the number of non-dyskinetic patients.

2. Type II: Seven patients were consistently labeled non-dyskinetic by the senior neurologist in all the three rank sets. These patients were uniformly eliminated from the original rank sets of all the neurologists and the remaining 28 patients were re-ranked keeping the order unchanged. UDysRS and SVS rankings were also modified accordingly.

Study I: Validation of ranking protocol - intra- and inter-neurologist agreement

Step (a) of the ranking protocol was based on the clinical definition of dyskinesia and not a specific rating scale. The ranking protocol was developed to compare the discrete neurologists' ratings in Step (a) to the continuous SVS score. By evaluating the intra- and inter-neurologist consistency in using the ranking protocol, the validity of the protocol can be determined. A high intra- and inter-neurologists consistency indicates that the protocol, based on clinical definition of dyskinesia, can be used to rank severity of dyskinesia by neurologists and in turn can be used to evaluate the utility of SVS. Independent analyses were performed to observe intra- and inter-neurologist agreement on the Type I and Type II ranking datasets. Kendall's tau-b correlation coefficient was computed pairwise between the four neurologists in each type to evaluate the inter-neurologist agreement. Intra-class coefficient for each neurologist across Set 1r, Set 2r, and Set 3r was computed. [17].

Study II: SVS Utility

The goal of this study was to assess if the SVS was a suitable score to quantify dyskinesia in clinical settings. Hence, we compared it to the UDysRS which was used as a gold standard by computing the Kendall's Tau-b correlation coefficient between SVS rankings and the UDysRS rankings. The SVS was also compared to neurologist's performance by computing the Kendall's Tau-b correlation coefficient between the SVS rankings and the neurologists' rankings. The statistical analysis was performed on Type I and Type II rankings of Set 1r, Set 2r and Set 3r. A good correlation would indicate that the SVS can quantify dyskinesia as well as neurologists and can complement UDysRS scores by providing an objective dimension to it.

Study III: Effect of ratings vs. rankings

We propose that ranking the severity of dyskinesia within each rating category of mild, moderate or severe dyskinesia assists in quantifying the differences between patients at a finer level. It is easier for neurologists to use a discrete five point rating scale which allows them to assign more than one patient in a single category than to use a ranking system which compels them to observe differences in severity more closely in order to assign individual ranks to each patient. If SVS was thresholded to represent a discrete scale, we wanted to compare the discrete SVS ratings to the neurologists' discrete ratings in Set 1R, Set 2R and Set 3R. Based on the a-priori information that approximately an equal number of patients were present in each rating category, the SVS was thresholded into four levels severity – absent, mild, moderate and severe. Kendall Tau-b correlation coefficient was computed between these SVS ratings and the neurologists ratings. These correlation coefficients were

compared to the correlation coefficients obtained in Study II which compares SVS rankings to neurologists' rankings. We would expect a higher correlation using the ratings than the rankings. In rankings, we compare the ordering of patients based on severity, whereas in ratings, groups of patients, irrespective of their ordering within the group, are compared. Our severity score SVS, in association with the ranking technique, can thus be used to complement rating scales to observe differences in dyskinesia severity in large patient databases.

Longitudinal evaluation of the SVS

SVS computed on three video segments of the same patient video was plotted and analyzed qualitatively. We could not perform statistical analysis on our data due to a small sample size of five subjects who had good video data for at least 30–45s. The goal was to observe if 10s of video data was sufficient to quantify dyskinesia severity using our technique.

RESULTS

Study I: Validation of ranking protocol - intra- and inter-neurologist agreement

Intra-neurologist agreement—Intra-class correlation coefficient (ICC) was computed only for Type I rankings for each neurologist as this measure was used to study intra-neurologist consistency. Hence Type II rankings, which were developed mainly for inter-rater studies, would not apply in this experiment. High ICC values were observed as follows: N1 – 0.9525; N2 – 0.948; N3 – 0.9496; N4 – 0.9928 ($p \leq 0.0001$).

A more detailed analysis was performed by computing pairwise Kendall tau-b values between the three sets of rankings for each neurologist and between the neurologists' rankings and the UDysRS rankings. Table 1 indicates that neurologist N1 and N4 are more consistent across the three sets of rankings compared to neurologist N2 and N3 which was also observed from the ICCs.

Inter-neurologist agreement—Though the neurologists independently ranked the videos based on the clinical definition of dyskinesia, high inter-neurologist agreement was observed between them. Type I and Type II rankings showed similar trends of agreement as seen in Figures 1 and 4. Kendall tau-b values ranged from 0.7754 to 0.8746 for Type I rankings, and from 0.6931 to 0.8095 for Type II rankings. All tau values were statistically significant with $p \leq 0.0001$.

Study II: SVS Utility

Figure 2 shows that Type I rankings across the three ranking sets exhibited higher Kendall tau-b values than Type II rankings. The tau values of the UDysRS rankings vs. SVS rankings and the neurologist's rankings vs. SVS rankings were comparable and indicated a moderate utility of the SVS as a dyskinesia quantifying score. The tau values ranged from 0.657166 to 0.7563 for Type I rankings, and from 0.5561 to 0.7143 for Type II rankings. All tau values were statistically significant with $p \leq 0.0001$.

Study III: Effect of ratings vs. rankings

Figure 2 shows that Kendall tau-b coefficients ranged from 0.6833 to 0.8519 for SVS ratings vs. neurologists' ratings which are higher when compared to the tau-b coefficients obtained in Study II while comparing rankings. These higher values indicate a more global agreement between the neurologists and the SVS.

Longitudinal evaluation of SVS

Figure 3 shows that the SVS computed on three different video segments of the same patient was reasonably close and the order of the patient severity remained the same when comparing SVS using the three different segments. Comparing SVS using different set of segments produced no overlap in the SVS values, thus maintaining the order of severity of the patients. For example: If Segment 1 for Patients 1, 2, and 3; Segment 2 for Patient 4; Segment 3 for Patient 5 were compared, the order of severity would remain unchanged.

DISCUSSION

To the best of our knowledge, this is the first study to report a video-based objective measure of dyskinesia using a single severity score. This score correlated closely with both the UDysRS obtained from viewing the entire clip and with clinicians ranking the 10 second clips used to obtain the the SVS. Although the SVS will not likely replace clinical rating scales such as the UDysRS, it has the advantage of being objective, automated and continuous. This makes it particularly appealing for small pilot studies screening agents for potential anti-dyskinesia effects.

Previous studies on quantifying dyskinesia [9,10,11 and 12] have used various devices such as accelerometers or gyroscopes. These record various metrics of movement instead of a single score. The advantages of using our quantifying technique include the widespread availability of video recording equipment, its ease of use in clinical settings and the capacity to share such recordings among colleagues immediately or as part of a longitudinal study.

We felt that it was important to demonstrate that our measure of dyskinesia correlated with clinical evaluations. The ranking protocol was designed based on the ratings assigned using the clinical definition of dyskinesia as opposed to a particular rating scale such as the CDRS or AIMS. Our results from Study I indicate that ranking protocol was valid and produced high inter- and intra-neurologists' consistency while using it. Our results from Study II, based on amplitude and directions of movement, indicate that the SVS moderately correlates with the UDysRS and the neurologists' rankings. The lower correlation coefficients in Type II rankings can be attributed to two possibilities: (a) reduction in the sample size from 35 to 28 video sequences and (b) the number of dyskinetic patients removed from each set was based on the observation of only the most senior neurologist. Amplitude and directions of movement are necessary but not sufficient parameters to completely quantify dyskinesia.

But in Study II, the ranking or the ordering of the patients within each group was compared thus requiring a more stringent quantifying technique which is capable of capturing the subtler difference between patients within that group. Hence we believe that using a ranking technique can initiate the development of better quantifying measures. Two patients can be

assigned the same category of severity using a rating scale, but not necessarily the same rank. This variation can occur as the neurologist is not restricted to the standard rating scale definition and may be inclined to assess the motor disability of one patient as more severe than the other, but not significant enough to assign the next higher rating on the scale. We do not reject the possibility of two patients receiving the same severity score and hence the same rank, but such an occurrence might be uncommon as SVS is a continuous variable.

We also observed in a small sample of patients, that 10s of video data is sufficient to quantify severity of dyskinesia using SVS. This longitudinal video segment study was just a proof of principle which can be further validated using longer video sequences dedicatedly obtained for improving and validating SVS. In this direction, further work is being done to improve and to automate the video tracking on longer video sequences and to include velocity of movement, anatomical distribution, and contributions from dystonia and chorea into the SVS. We believe that these factors will increase the correlation between the SVS and the neurologists' rankings. Such a severity score can be applied to other assessments tasks similar to the activities of daily living and used to quantify longer video sequences. Dyskinesia exhibits diurnal variations and our study is based on 10s video sequences of a specific task. Hence even though our semi-automatic technique is comparable to standardized rating scales, which take such diurnal variations into account, it should be views as a complement not a replacement for them.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank Dr. Christopher Goetz and Dr. Glenn Stebbins for providing the 35 video sequences, the UDysRS scores and the required informed consent forms to perform this study and their valuable comments and suggestions on reviewing the manuscript. This research work was supported by a Vanderbilt University Discovery Grant and NIH Grant R01EB006136.

REFERENCES

1. Rosal-Greif VLF. Drug induced dyskinesias. *Am J Nurs*. 1982; 82:66–69. [PubMed: 6119904]
2. Goetz CG, Nutt JG, Stebbins GT. The Unified Dyskinesia Rating Scale: Presentation and clinimetric profile. *Mov Disord*. 2008; 23(16):2398–2403. [PubMed: 19025759]
3. Guy, W. AIMS: ECDEU assessment manual for psychopharmacology. US Washington, DC: Government Printing Office; 1976.
4. Parkinson Study group. Evaluation of dyskinesia in a pilot randomized placebo-controlled trial of ramaecide in advanced Parkinson's disease. *Neurology*. 2001; 58:1660–1668.
5. Goetz CG, Stebbins GT, Shale HM, Lang AE, Chernik DA, Chmura TA, et al. Utility of an objective dyskinesia rating scale for Parkinson's disease inter- and intra-rater reliability assessment. *Mov Disord*. 1994; 9:390–394. [PubMed: 7969204]
6. Katzenschlager R, Head J, Schrag A, Ben-Shlomo Y, Evans A, Lees AJ. Parkinson's disease research group of the united Kingdom, Fourteen year final report of the randomized PDRGUK trial comparing three initial treatments in PD. *Neurology*. 2008; 71:474–480. [PubMed: 18579806]
7. Hagell P, Widner H. Clinical Rating of Dyskinesias in Parkinson's disease: Use and Reliability of a New Rating Scale. *Mov Disord*. 1999; 14(3):448–455. [PubMed: 10348468]

8. Hobart JC, Cano SJ, Zajicek JP, Thompson AJ. Rating scales as outcome measures for clinical trials in neurology: problems, solutions, and recommendations. *Lancet*. 2007; 6(12):1094–1105.
9. Keijsers NLW, Horstink MWIM, Gielen SCAM. Automatic Assessment of Levodopa-Induced Dyskinesia in Daily Life by Neural Networks. *Mov Disord*. 2003; 18(1):70–80. [PubMed: 12518302]
10. Hoff JI, Wagemans EAH, Van Hilten JJ. Accelerometric Assessment of Levodopa-Induced Dyskinesias in Parkinson's Disease. *Mov Disord*. 2001; 16(1):58–61. [PubMed: 11215593]
11. Patel S, Sherrill DM, Hughes R, Hester T, Huggins N, Lie-Nemeth T, et al. Analysis of the Severity of Dyskinesia in Patients with Parkinson's Disease via Wearable Sensors. *Proceedings of the International Workshop on Wearable and Implantable Body Sensor Networks*. 2006:123–126.
12. Liu X, Carroll CB, Wang S, Zajicek J, Bain PG. Quantifying Drug-induced Dyskinesias in the Arms Using Digitized Spiral-Drawing Tasks. *J Neurosci Methods*. 2005; 144(1):47–52. [PubMed: 15848238]
13. Rao AS, Bodenheimer RE, Davis TL, Li R, Voight C, Dawant BM. Quantifying Drug Induced Dyskinesia in Parkinson's disease Patients Using Standardized Videos. *Proceedings of the 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. 2008:1769–1772.
14. Rohde GK, Aldroubi A, Dawant BM. The Adaptive Bases Algorithm for Intensity-Based Non-rigid Image Registration. *IEEE Trans Med Imaging*. 2003; 22(11):1470–1479. [PubMed: 14606680]
15. Jolliffe, IT. *Principal Components Analysis*. New York: Springer-Verlag; 1986.
16. Kendall, MG. *Rank Correlation Methods*. London: Charles Griffin and Company Limited; 1948. p. 25-36.
17. Shrout PE, Fleiss JL. Intraclass correlations: Uses in assessing rater reliability. *Psychol Bull*. 1979; 86:420–428. [PubMed: 18839484]

Figure 1.

Scatter plots showing inter-neurologist Type I ranking correlations. (First row) N1 vs. N2, N3 and N4; (second row) N2 vs. N3 and N4; (third row) N3 vs. N4 Kendall tau-b coefficient is indicated in parenthesis.

Figure 2. Scatter plots showing correlation between SVS and Type I neurologist' rankings (first row); SVS and Type II neurologists' rankings (second row); SVS and neurologists ratings (third row); SVS and UDysRS rankings (fourth row). Kendall tau-b coefficient is indicated in parenthesis.

Figure 3.

SVS computed on three different video segments of five patients. Patient 3 did not have good Segment 2 data indicated by the missing green bar in Segment 2.