A comparison of clinical and objective measures of freezing of gait in Parkinson’s disease

Tiffany R. Morris a, Catherine Cho b, Valentina Dilda a, James M. Shine c, Sharon L. Naismith c, Simon J.G. Lewis c, Steven T. Moore a, b, * 

a Human Aerospace Laboratory, Department of Neurology, Mount Sinai School of Medicine, Box 1052, 1 Gustave L Levy Place, NY 10029, USA 
b Robert and John M. Bendheim Parkinson and Movement Disorders Center, Department of Neurology, Mount Sinai School of Medicine, NY 10029, USA 
c Parkinson’s Disease Research Clinic, Brain and Mind Research Institute, University of Sydney, Sydney, Australia

ARTICLE INFO

Article history:
Received 8 December 2011
Received in revised form 27 February 2012
Accepted 1 March 2012

Keywords:
Gold standard 
Accelerometry 
FOG 
Timed up-and-go task

ABSTRACT

Freezing of gait, a paroxysmal motor block, is common in the latter stages of Parkinson’s disease. The current ‘gold standard’ of assessing the severity of freezing is based on clinical identification (by up to 3 raters) of the number of episodes from video. The aims of this study were to systematically assess this ‘gold standard’ across multiple Parkinson’s disease centers, and to compare these clinical ratings with objective measures derived from lower limb acceleration data. Video recordings were acquired during a timed up-and-go task from 10 Parkinson’s disease patients (with a clinical history of freezing) in the ‘off’ state. Patients were instrumented with accelerometers on the lateral aspect of each shank. Ten experienced clinicians were recruited from four Parkinson’s disease centers to independently assess the videos for number and duration of freezing events. The reliability of clinical video assessment for number of freezing events was moderate (intraclass correlation coefficient 0.63). Percent time frozen (cumulative duration of freezing episodes/total duration of the walking task) demonstrated stronger agreement between raters (0.73). Agreement of accelerometry-derived measures of freezing severity with mean clinician ratings was strong for number of episodes (0.78) and very strong for percent time frozen (0.93). The results demonstrate the viability of objective measures of freezing, and that percent time frozen is a more reliable metric of severity than number of freezing events for both clinical and objective measures. The large variability between clinicians suggests that caution should be used when comparing subjective ratings across centers.

© 2012 Elsevier Ltd. All rights reserved.

1. Introduction

Freezing of gait (FOG), a transient block of movement triggered by gait initiation, turning, or obstacle avoidance, commonly affects patients in the latter stages of Parkinson’s disease (PD) [1]. The pathophysiological mechanisms underlying FOG remain poorly understood [2], and response to current treatments (including dopamine replacement therapy [3]) is at best limited and more often ineffective [4]. FOG represents a common cause of falls in PD [3] and significantly impairs quality of life [5].

Clinical management of this disabling symptom is limited in large part by the difficult nature of assessing FOG severity. Given its paroxysmal nature, patients may appear FOG-free in the clinical setting, although evaluation in the ‘off’ state increases the likelihood of observing freeze events [6]. Subjective measures of FOG have predominated in both clinical and research settings. The Unified Parkinson’s Disease Rating Scale part 14 ‘Freezing When Walking’ (UPDRS 14) [7] rates patients on a scale from 0 (none) to 4 (frequent falls from freezing) based on clinical history, and has traditionally been considered a reliable questionnaire-based indicator of FOG [8]. More recently, the Freezing of Gait Questionnaire (FOG-Q) [8] and the new FOG-Q (NFOG-Q) [9] have been proposed as tools to identify freezing behavior and assess the efficacy of interventions (FOG-Q question 3, Do you feel that your feet get glued to the floor while walking, making a turn or when trying to initiate walking (freezing)?, distinguished ‘freezers’ at least as well as UPDRS 14 [8]). However, validation of these FOG questionnaires was subjective, relying on patient and caregiver responses and clinician-mediated questionnaires [8,9]. In our recent study of a cohort of PD patients with self-reported FOG symptoms [10], neither the FOG-Q nor NFOG-Q score correlated with the frequency

Please cite this article in press as: Morris TR, et al., A comparison of clinical and objective measures of freezing of gait in Parkinson’s disease, Parkinsonism and Related Disorders (2012), doi:10.1016/j.parkreldis.2012.03.001
and duration of freezing episodes observed in the ‘off’ state. Thus, the utility of questionnaire-based FOG assessments may well be limited to the identification of freezing behavior, of which only a single question (UPDRS 14, FOQ-G 3) appears necessary.

In the past decade a de facto ‘gold standard’ of FOG assessment has emerged; clinical evaluation of video recordings of ambulating patients utilizing one [11–15], two [10,15–19], or three raters [20–23]. Less commonly identified as neurologists specializing in movement disorders [14,17,21] or physiotherapists [23], the majority of studies generically classify observers along the lines of having ‘extensive experience in FOG assessment’ [22] or more simply as ‘independent observers’ [18]. Agreement between raters is typically high, with intraclass correlation coefficients (ICC) [24] greater than 0.8 [10,18,20,22], often achieved by consensus [15–17,19] or averaging of observations [20,22,23]. Although not stated explicitly, raters are likely to be from the same or collaborating institutions, authors on the study, or both (the latter is certainly the case for our group [10,14] and that of Ziegler [23]).

The number of FOG episodes observed was the principal measure in almost all of the studies cited above. The primary aim of the current study was to test clinical video evaluation as a ‘gold standard’ of FOG assessment across different institutions with a larger cohort of observers (N = 10), in particular exploring the reliability (ICC) between truly independent, experienced raters. Surprisingly, no such systematic test of what is now explicitly labeled the ‘gold standard’ [15,17] has been conducted to date.

The secondary aim of this study was to compare a novel objective approach to the assessment of freezing frequency and severity, based on lower limb accelerometer (proposed by Moore et al. [14]), with the current ‘gold standard’ of video assessment. Validation of objective ambulatory assessment of FOG may improve clinical management of patients with freezing of gait.

2. Methods

2.1. Recruitment

Ten patients (6 male, 4 female) who were attending the Parkinson’s Disease Research Clinic at the Brain and Mind Research Institute were identified for this study by self-reported freezing behavior. All patients satisfied UPDRS Parkinson’s disease criteria [25], had a Mini-Mental State Examination (MMSE) [26] score of ≥24 and were deemed unlikely to have dementia or major depression according to DSM-IV criteria by consensus rating of a neurologist (SJGL) and a neuropsychologist (SLN). The study was approved by The University of Sydney Human Research and Ethics Committee and written informed consent was obtained.

2.2. Clinical evaluation and questionnaires

Patients were assessed in the practically-debrieved ‘off’ state following overnight withdrawal of dopaminergic therapy. Two patients were being treated with deep brain stimulation (DBS), which was turned off for 60 min prior to testing. Patient characteristics were as follows; mean age 67.7 (SD 6.6), disease duration 11.2 years [SD 6.7], Hoehn and Yahr stage [27] 2.5 (SD 0.3), UPDRS Section III 38.4 (SD 10.7), Parkinson’s Disease Questionnaire (PDQ) 39 [28] 64.0 [SD 23.9], and NFOG-Q 17.7 [SD 5.9]. None of the patients described any increase in freezing behavior following the administration of their usual dopaminergic therapy.

2.3. Locomotor task and data acquisition

Patients performed timed up-and-go (TUG) tasks to provoke FOG on a standardized course (Fig. 1), as described previously [10]. Walking trials were recorded on a digital video camera from a consistent vantage point for later analysis [10], and each video showed a complete TUG trial starting and ending in the seated position. Fourteen TUG tests were performed with a total duration of 9 min 37 s (four of the ten subjects who performed the TUG task quickly, or who exhibited minimal FOG, were asked to perform a second trial). The overall aim of the investigators was to obtain a sequence of videos depicting a wide range of FOG severity, whilst limiting the time required to clinically rate the collected videos to 45 min or less (several viewings were required for each clip).

Simultaneous acceleration data was acquired from the lower limbs during the TUG trials described above. Prior to testing patients were instrumented with small (38 × 53 × 21 mm; 30 g) inertial measurement units (MTX, Xorn, Enschede, The Netherlands) attached to the lateral aspect of each shank (just above the ankle). Each unit acquired longitudinal shank linear acceleration at a sample rate of 50 Hz, which was transmitted wirelessly to a computer for analysis. Synchronization of the video and accelerometer recordings was performed prior to data collection by alignment of the video camera and data acquisition computer clocks.

2.4. Clinical assessment of FOG

Ten clinicians from four Parkinson’s disease centers, experienced in evaluation of freezing, were recruited to independently evaluate the videos for FOG:

1. Bendheim Parkinson and Movement Disorders Center, Mount Sinai School of Medicine, New York, NY (four neurologists),
2. National Parkinson Foundation Center of Excellence, Beth Israel Medical Center, New York, NY (three neurologists),
3. James J. Peters VA Medical Center, Bronx, NY (one neurologist),
4. Parkinson’s Disease Research Clinic, University of Sydney, Australia (one neurologist and one clinician).

The frequency and duration of freezing episodes were determined for each video utilizing a FOG tagging program developed by two non-rater investigators (TRM and STM). Each clinician performed the FOG assessment independently, in the presence of an investigator (TRM or STM) who explained the operation of the software and provided technical assistance as required. The observers were instructed to use their best clinical judgment to identify the onset and termination of freezing episodes, and that the investigators did not offer any input regarding the assessment of the video recordings. A total of 18 videos (each depicting an individual TUG trial) were presented to each rater in a pseudorandom order, with no two observers having the same sequence. Each comprised of the 14 unique TUG trials from 10 patients to assess inter-rater (across the 10 observers) reliability, plus four repeat videos to assess intra-rater (within each observer) reliability. Raters were not informed that some patients had more than one video trial or that four of the videos were presented twice.

Clinicians were instructed to first watch each TUG video in its entirety. On subsequent viewings, raters were required to tag the onset of a freeze by pressing the ‘T’ key and to hold down the key throughout the duration of each event. Video editing tools enabled the ends of a horizontal bar representing the duration of each tagged freeze to be dragged backwards or forwards in time (which also moved the video by a corresponding amount) to log the onset and offset of each FOG event (Fig. 1). To facilitate accurate assessment of freeze duration, the raters were instructed to repeatedly view the video footage and adjust onset/offset points until entirely satisfied, then press a ‘save’ button. At this point the software logged the tag data to disk and presented the next video for assessment.

The program saved the clinical ratings as a binary signal at a sample rate equivalent to the video frame rate (30 Hz), with a baseline of zero (no freeze) and a value of 1 indicating a freeze event (Fig. 2A–C). From these binary signals two metrics were computed for each video trial to quantify freezing of gait. Number of FOG episodes was defined as the number of separate freeze tags, which could take the form of transitions (Fig. 2A–C) from baseline to 1 to baseline (freezing during walking or turning), 1 to baseline (start hesitation), or baseline to 1 (target hesitation upon freezing of gait). In order to minimize inadvertent tagging due to movement, any two FOG tags separated by less than 1 s were consolidated as a single freeze event. The percent time frozen was computed by summing the length of all FOG tags and dividing by the total duration of the walking task (from initiation of the first step to the conclusion of the final step).

2.5. Objective assessment of FOG

A member of our research team (STM) has previously described a technique for objective identification of freezing episodes based on the frequency characteristics of vertical shank acceleration [14]. The algorithm is predicated on the assumption that freezing of gait entails an increase in high frequency leg movement (‘trembling’) in the relative absence of locomotor activity [14]. An index of freezing of gait (iFOG) was defined as the ratio between (i) the cumulative power of the power spectra within a temporal window (set to 10 s in this study) in a ‘freeze’ band (3–8 Hz) and (ii) the cumulative power in the ‘locomotor’ band. This index was calculated within a 10 s window centered at each point in time (i.e., every 20 ms) resulting in a continuous waveform, and a composite iFOG signal was produced by combining data from the left and right shank (Fig. 2D). A discrete freeze event was defined as a contiguous period of time in which iFOG was greater than 2; that is, the power in the ‘freeze’ band was at least twice that in the ‘locomotor’ band. A binary signal analogous to that produced by the video tagging program, was created and down-sampled to 30 Hz to allow comparison with the clinical video scoring (Fig. 2E).

2.6. Statistical analysis

The reliability of the quantification of the number of FOG episodes and percent time frozen was calculated between the 10 clinical observers, within each observer.
for the four repeat assessments, for each possible pairing of observers, between the mean rating of all 10 observers and the objective accelerometer-derived (iFOG) measures, and between each individual observer and iFOG, using the intraclass correlation coefficient (ICC) [24]. In accordance with Schafsma [22], we used the following classification of ICC power: <0.2 negligible, 0.2 ≤ 0.4 weak, 0.4 ≤ 0.7 moderate, 0.7 ≤ 0.9 strong, and >0.9 very strong. In addition, the mean and 95% confidence interval of the number of FOG events and percent time frozen from the 10 raters was calculated for each of the 14 video trials.

3. Results

There was a range of freezing severity observed in the 14 videos by the 10 clinicians; mean number of FOG events reported in a single trial (Fig. 3A) varied from 0 (CI 0) to 6 (CI 3.0); percent time frozen varied from 0 (CI 0) to 79.8 (CI 6.0) (Fig. 3B). The reliability of clinical ratings between the 10 observers was moderate (ICC = 0.63) for number of FOG, but stronger (ICC = 0.73) for percent time frozen. Intra-rater reliability was remarkably low for number of FOG (0.44 (CI 0.18)), but was strong (0.71 (CI 0.19)) for percent time frozen. A comparison between all possible pairs of raters revealed a wide range of agreement, from negligible to very strong, for both number of FOG (mean 0.45 [CI 0.07]) and percent time frozen (mean 0.76 [CI 0.05]). There was no significant difference (P > 0.05) in ICC for rater pairs from different PD centers (N = 35) compared to pairings from within the same institution (N = 10) for number of FOG (0.42 [CI 0.18] versus 0.46 [CI 0.08]) and percent time frozen (0.80 [CI 0.08] versus 0.75 [CI 0.05]).

Reliability between each individual rater and the accelerometry-derived (iFOG) measures ranged from weak (ICC = 0.37) to strong (ICC = 0.83) for number of FOG, with an average of 0.62 (CI 0.07). Agreement between individual observers and iFOG was considerably higher for percent time frozen, ranging from moderate (ICC = 0.68) to very strong (ICC = 0.96), averaging 0.83 (CI 0.06). The average score of all 10 clinicians provides a more robust estimate of FOG severity than individual ratings, and there was strong agreement between the number of FOG events detected by accelerometry and the mean of all 10 clinicians (ICC = 0.78) (Fig. 3A), and an even stronger correlation (ICC = 0.93) between objective assessment of percent time frozen and the average of the 10 raters (Fig. 3B).

To evaluate the robustness of objective FOG scoring we analyzed whether the choice of window size and iFOG threshold affected the reliability of accelerometry-derived measures when compared to
the average of the 10 clinical raters. The width of the temporal window used in the calculation of the FOG index represents a tradeoff between sensitivity to short FOG events (smaller window) and a sufficient number of data points for a robust power analysis (larger window). In a previous study, in which leg acceleration was sampled at 100 Hz, we found a window of 6 s to be sensitive to freezing episodes as short as 2 s [14]. In the current study, in which data acquisition was lower at 50 Hz, we utilized a 10 s window for the frequency analysis. We calculated the ICC between iFOG and raters at window sizes ranging from 2 to 29 s (Fig. 4A); ICC was robust for windows above 5 s, with the optimum range for number of FOG being 8–29 s, and 6–15 s for percent time frozen. In a similar manner we calculated ICC for iFOG thresholds (ratio of 'freeze' power to 'locomotor' power) ranging from 1 to 6 (with window size set to 10 s); the reliability of objective accelerometry scoring was maintained over the entire range for both number of FOG and percent time frozen (Fig. 4B).

4. Discussion

The results of this study demonstrate that the reliability of clinical video assessment for number of FOG events, the current ‘gold standard’ [15,17], was not robust (ICC = 0.63) either within or across multiple Parkinson’s disease centers; moreover, intra-rater reliability was remarkably low (mean ICC 0.44). A new metric proposed by our group, percent time frozen, demonstrated stronger inter- (ICC = 0.73) and intra-rater (mean ICC = 0.71) agreement. Objective measurement of number of FOG episodes derived from lower limb accelerometry (developed by Moore et al. [14]) was in strong agreement with the mean scoring of 10 raters (ICC = 0.78). Objective accelerometer scoring of percent time frozen was even more strongly correlated with the clinician mean rating (ICC = 0.93).

The agreement between raters for scoring of the number of FOG events in the current study was considerably lower than that reported in previous studies utilizing video assessment. Careful attention is required in defining the onset and termination of FOG, and it is perhaps not surprising that small research groups, in which consensus is likely reached on the definition of FOG prior to rating [15–17,19], achieved higher agreement than a larger cohort of independent observers. Although this approach may be valid within a study group in which consistency is likely to be maintained (for example, pre- and post-intervention), caution should be used

![Fig. 2. A–C Binary waveform (1 indicates freezing) from three observers rating the same video clip during a TUG task of 25-s duration. D The iFOG waveform from the same trial, derived from the relative power of shank acceleration in the 'freeze' (3–8 Hz) and locomotor' (0.5–3 Hz) frequency bands. In this study any periods in which the iFOG signal was greater than 2 was labeled a freezing event. E Binary waveform (1 indicates freezing) derived from the iFOG waveform.]
Percent time frozen proved a more reliable metric of FOG severity than simply counting the number of FOG events for both clinical and objective measures. The basis for this finding is readily apparent when considering the results from three raters scoring the same video in Fig. 2, in which raters 7, 9 and 8 observed one, two and three distinct FOG events with percent time frozen of 37.0, 33.0, and 29.6 (the objective iFOG approach scored a single event with percent time frozen of 32.3). Whether the patient was seen to have experienced multiple sequential FOG or one long event is less critical when percent time frozen is calculated; calculation of the relative duration of freezing minimizes differences in inter-rater preference for registering multiple or concatenated FOG episodes.

The results of this study provide strong justification for objective FOG monitoring based on lower-body acceleration. The finding that autonomous freeze identification from shank accelerometry [14] was in strong agreement with the mean ratings of 10 independent experienced clinicians across four Parkinson’s disease centers, for both number of FOG events (ICC = 0.78) and percent time frozen (ICC = 0.93), demonstrates the viability of this approach and suggests that high frequency (3–8 Hz) leg movement (whether visible or sub-clinical) is present in the majority of FOG episodes.

Based on the results of this study we provide a number of practical recommendations for clinical FOG assessment; 1) at least two observers should score patient videos; 2) the criteria for identifying FOG onset and offset should be clearly reported; and 3) percent time frozen should be utilized as a metric in conjunction with number of FOG episodes. In addition, lower-body accelerometry should be considered as an objective alternative to clinical FOG assessment. The growing presence of accelerometers in mobile computing platforms (such as smart phones) may provide a means to implement this methodology in the clinic [29,30], although much work remains to be done to validate such an approach.

Financial support

Dr. Moore is supported by the National Space Biomedical Research Institute and NASA grant NNX09AL14G; Dr. Cho is supported by NIH grant 5R01DC007658-05; Dr. Naimsmith is supported by an NHMRC Career Development Award No. 1008117; Dr. Lewis is supported by an NHMRC Practitioner Fellowship and a University of Sydney Rolf Edgar Lake Postdoctoral Fellowship; there are no financial disclosures for Drs. Shine, Morris or Dilda to make.

Acknowledgements

The authors gratefully acknowledge the patients and clinicians who participated in this study.

References


