Journal of Parkinson's Disease xx (2022) x–xx
DOI 10.3233/JPD-223493
IOS Press

Research Report

An Evaluation of KELVIN, an Artificial Intelligence Platform, as an Objective Assessment of the MDS UPDRS Part III

Krista Sibley[a], Christine Girges[a], Joseph Candelario[a], Catherine Milabo[a], Maricel Salazar[a], John Onil Esperida[a], Yuriy Dushin[b], Patricia Limousin[a] and Thomas Foltynie[a,*]

[a]*Department of Clinical & Movement Neurosciences, UCL Institute of Neurology, National Hospital for Neurology and Neurosurgery, London, UK*

[b]*Machine Medicine Technologies, London, UK*

Accepted 31 August 2022

Abstract.

Background: Parkinson's disease severity is typically measured using the Movement Disorder Society Unified Parkinson's disease rating scale (MDS-UPDRS). While training for this scale exists, users may vary in how they score a patient with the consequence of intra-rater and inter-rater variability.

Objective: In this study we explored the consistency of an artificial intelligence platform compared with traditional clinical scoring in the assessment of motor severity in PD.

Methods: Twenty-two PD patients underwent simultaneous MDS-UPDRS scoring by two experienced MDS-UPDRS raters and the two sets of accompanying video footage were also scored by an artificial intelligence video analysis platform known as KELVIN.

Results: KELVIN was able to produce a summary score for 7 MDS-UPDRS part 3 items with good inter-rater reliability (Intraclass Correlation Coefficient (ICC) 0.80 in the OFF-medication state, ICC 0.73 in the ON-medication state). Clinician scores had exceptionally high levels of inter-rater reliability in both the OFF (0.99) and ON (0.94) medication conditions (possibly reflecting the highly experienced team). There was an ICC of 0.84 in the OFF-medication state and 0.31 in the ON-medication state between the mean Clinician and mean Kelvin scores for the equivalent 7 motor items, possibly due to dyskinesia impacting on the KELVIN scores.

Conclusion: We conclude that KELVIN may prove useful in the capture and scoring of multiple items of MDS-UPDRS part 3 with levels of consistency not far short of that achieved by experienced MDS-UPDRS clinical raters, and is worthy of further investigation.

Keywords: Artificial intelligence, clinical trials, digital measures, Parkinson's disease, remote monitoring

INTRODUCTION

The quantitative assessment of Parkinson's disease (PD) is important for the study of the natural history of PD progression, as well as for the assessment of conventional interventions, for example response to dopaminergic therapy or neurosurgical interventions such as deep brain stimulation (DBS) and also for evaluating the impact of novel or experimental interventions in clinical trials. The Movement Disorder Society Unified Parkinson's Disease Rating Scale Part 3 (MDS-UPDRS part 3) is the most commonly used standard assessment tool for measuring

[*]Correspondence to: Professor T. Foltynie, PhD, Box 146, National Hospital for Neurology & Neurosurgery, Queen Square, London. WC1N 3BG, UK. Tel.: +44 0 203 448 8726; E-mail: T.Foltynie@ucl.ac.uk.; ORCID: 0000-0003-0752-1813

the motor signs of Parkinson's disease (PD) [1]. It consists of 18 items quantifying rigidity, bradykinesia, tremor, and axial signs (namely facial expression, speech, posture, gait, and balance). Signs are rated on a 5-point scale (0–4) with separate scores for an individual's left and right hemibody. The MDS-UPDRS is frequently used as an objective primary outcome measure for clinical trials of novel and experimental PD neurotherapeutics [2].

A standardized training program exists for administering the MDS-UPDRS (see https://www.movementdisorders.org/MDS/Education/Rating-Scales/Training-Programs.htm) with a teaching tape that has been shown to improve MDS-UPDRS ratings and inter-rater reliability [3]. In this context, previous literature has quantified the intra-rater and inter-rater variability of the MDS-UPDRS, with inter-rater variability showing good but not perfect Intraclass Correlation Coefficient (ICC) [4] scores for the sum of the MDS-UPDRS part 3 between 0.65–0.91 [5, 6]. Studies measuring intra-rater variability generally show excellent reliability for the sum of the MDS-UPDRS, for experienced raters in neurology and movement disorder specialists when measured between 1–8 weeks apart (ICCs between 0.90–0.91) [5, 7, 8].

For individual items of the MDS-UPDRS, studies measuring inter-rater reliability have shown variable results with kappa scores showing anything between substantial and near perfect agreement (ICC 0.63–0.92) for items of bradykinesia [5, 6, 9, 10], moderate to near perfect agreement for gait and posture related items (ICC 0.49–0.93) [5, 10], fair to near perfect agreement for speech and facial expression (ICC 0.22–0.83) [5, 10], and fair to near perfect agreement for tremor items (ICC 0.31–0.90) [5]. Studies measuring the intra-rater agreement across specific items of the MDS-UPDRS also show mixed results with tremor showing moderate to near perfect agreement (ICC 0.43–0.93), bradykinesia items showing moderate to near perfect agreement (ICC 0.59–0.90), rigidity demonstrating substantial agreement (ICC 0.61–69), posture and gait related items showing substantial to near perfect agreement (ICC 0.64–0.95), and speech and facial expression showing moderate to near perfect agreement (ICC 0.58–0.89) [5, 7, 8]. While generally good, this research indicates that measures of inter-rater and intra-rater agreement for items of the MDS-UPDRS are still potentially variable across studies.

The existence of this variability can have a major impact on the objective assessment of repeated measures of PD motor severity, which can lead to errors in the interpretation of natural history studies or when assessing the impact of therapeutic interventions. To help address this variability, several artificial intelligence (AI) tools are in development [11, 12]. Studies inputting data from wearable sensors and voice recordings into machine-learning techniques have demonstrated high levels of accuracy of automated systems to score items from the MDS-UPDRS such as bradykinesia [13–15] and tremor [14, 17–20]. However, wearable sensors are expensive and sometimes difficult to distribute, and exposure to multiple devices can be burdensome to patients.

Since the COVID19 pandemic, the emergence of video-conference platforms have been explored as an accessible and cost-effective way to assess the severity of motor PD, in the absence of face to face appointments [18]. Videos allow clinicians to complete the majority of items within the MDS-UPDRS, with the exception of rigidity and postural instability which both require a hands-on assessment of the patient. Given the cost-effective and accessible nature of video assessment of PD motor severity, they provide the scope for AI to be utilized as a potential alternative to traditional clinician rating of items such as bradykinesia which in theory might improve the consistency and objectivity of MDS-UPDRS assessments.

Rationale and aims

In this study, we assessed the extent to which an artificial intelligence platform KELVIN, in its development stage, compares in its assessment of specific items of PD motor severity compared to traditional clinical MDS-UPDRS part 3 scoring. PD patients were scored simultaneously by two clinicians and two accompanying KELVIN measurements. The purpose of this study was to compare the traditional clinical scores with the KELVIN scores, compare variation in scores within and between clinical observers, and the variation in scores calculated within the KELVIN automated platform.

METHODS

Participants

All participants attended the Unit of Functional Neurosurgery, National Hospital for Neurology & Neurosurgery, Queen Square, UK as part of their NHS care between 2020-2021. All patients had a diagnosis of PD and were undergoing a levodopa challenge test to assess their appropriateness for advanced therapies for PD. Participants were of Hoehn and Yahr stage 3 or less. Ethical approval for the capture of the Kelvin data was obtained from the National Hospital for Neurology and Neurosurgery Research Ethics Committee (19/YH/0421) and written informed consent was obtained from all the participants.

Assessments

Assessors comprised two experienced MDS-UPDRS raters (KS & CG) and four DBS nurse specialists (JC, CM, MS, JE). All assessors had extensive familiarity with the MDS-UPDRS and had reviewed the Movement Disorder Society training package for the motor section of the MDS-UPDRS [1] and passed the associated evaluation materials before any assessments were performed.

OFF-medication assessment

All participants attended the Unit of Functional Neurosurgery in the OFF state having stopped their conventional PD medications overnight (at least 12 h since their last dose). Participants underwent simultaneous clinical assessment of the motor severity of their PD by two assessors using the conventional MDS-UPDRS part 3, 18 item motor subscale (referred to as C18-UPDRS). Each patient was given the usual instructions to perform each item of the MDS-UPDRS part 3 by one of the assessors and each assessor then noted a clinical score for each MDS-UPDRS part 3 item without conferring. The MDS-UPDRS part 3 was conducted, and video recorded by one assessor using a tripod-mounted tablet with the KELVIN web-app and simultaneously by the second assessor using a separate tripod-mounted tablet with the KELVIN web-app, facilitating combined video capture alongside traditional clinical scoring of all the individual components of the subscale. Video recordings via tablets were necessarily placed at conveniently different distances and angles between patient and each tablet to maintain a full view of the patient without obstruction.

ON-medication assessment

Each participant took their usual L-dopa dose (or equivalent as Madopar dispersible), which typically took 1 h to take effect. Once a participant confirmed that their medications were starting to work in the usual way, participants underwent repeat evaluations.

Simultaneous video capture using KELVIN

The details of the KELVIN software have been previously described [19]. In short, video data captured by a consumer grade smartphone or tablet can be recorded, stored on the device, then later uploaded, re-accessed, and analyzed by the Kelvin-PD$^{\text{TM}}$ motor assessment software without need for participants wearing markers or any other wearable device. Video segments were saved within the app as individually catalogued files according to each sub-item of the MDS- UPDRS part 3. Videos were encrypted and coded prior to being stored in the KELVIN cloud system and were only accessible by permission granted to the clinical team under account access settings within the system. Users were required to login to their own personal account within KELVIN, in order to access videos, with a strong password policy enforced to ensure confidentiality.

KELVIN scoring of videos

Video segments for each of 7 items (See Table 1) of the MDS-UPDRS part 3 examinations were analyzed using a web-app version of KELVIN (https://KELVIN.machinemedicine.com/). The latest version of the KELVIN app automatically detects and defines the regions (time-periods) of interest (ROIs) during which the participant is performing the relevant movement; For example, finger-tapping videos would usually contain two ROIs, corresponding to the sections of the video in which the patient performed the action using their left and right hand. Signals were extracted from these KELVIN defined ROIs of the videos (Fig. 1).

Kelvin analytic processes

The Kelvin app uses the deep learning library OpenPose to extract 25 body and 21 hand key-point coordinates on each frame. OpenPose is a popular open-source library that constructs time-series signals based on the change of these key-points through time, and features were then extracted from these signals. For each of 7 MDS-UPDRS items, signals based on key-points relevant to the appropriate action were constructed (Table 1). A patient's signals were normalized using their estimated standing height,

Table 1
Signals constructed from key-points and used for feature extraction for each of the K7UPDRS items. Courtesy of Machine Medicine Technologies Ltd

MDS-UPDRS item	Time series signal
Finger Tapping	Euclidean distance between the thumb tip and the index finger tip.
Hand Movement	The area of the convex hull of the four finger tips and the palm.
Pronation Supination	The velocity of the angle between the little finger tip and the thumb tip.
Toe Tapping	The vertical distance between the big toe and the neck.
Leg Agility	The Euclidean distance between the knee and the neck.
Arise from Chair	The Euclidean distance between the nose and the midpoint of the two ankles and the Euclidean distance between the two wrists divided by the Euclidean distance between the shoulders.
Gait	The leg ratio difference, The vertical angle of the body, The horizontal angle of the ankles, the horizontal angle of the wrists, the horizontal distance between the heels.

with the exception of the pronation supination signal which was an angular measure and thus much less dependent on the distance between the patient and the camera.

Kelvin uses a peak detection algorithm [20] to identify local maxima (peaks) and minima (troughs), which typically correspond to the start and midpoint of a periodic action. For example, as the finger-tapping signal was based on the distance between thumb and index finger tip, a peak would correspond to the two fingers being maximally apart, and a trough would correspond to the two fingers touching.

For each of the five bradykinesia items, the relevant time-series signal captured the key characteristics of movement: frequency, amplitude, velocity, and smoothness of the actions. Patients with more severe impairment slow down earlier, execute actions less smoothly, with more rapid amplitude decrement. For the Arise from chair item, four features were extracted, intended to capture key characteristics of the examination listed in the MDS-UPDRS instructions. Patients with more severe impairment are slow to arise, need more than 1 attempt, and use the hands to push up from the armrests to get up from the chair [20]. For gait, step frequency (speed), two features relating to patients arm swing, and two features to capture roughness of walking, and variability in stride width, and a feature to measure postural control were combined to capture the MDS-UPDRS gait assessment [20].

Inter and intra-rater reliability of clinician scores

C18-UPDRS scores from Rater 1 and Rater 2 were calculated according to the MDS-UPDRS part 3 instructions. Scores for individual items were compared between assessors in subsequent analysis (inter rater reliability). To examine the extent to which the same experienced assessor has residual variability in their clinical ratings of the MDS-UPDRS part 3, each patient video was re-scored by Rater 1 at a second time point, blinded to previous scores. Sum C18 UPDRS part 3 and individual item scores were calculated at time 1 and time 2 (intra-rater reliability).

Inter-rater reliability of KELVIN

The KELVIN app automatically derived K7-UPDRS scores for each patient based on the 7 Kelvin rateable items (arising from chair, gait, and 5 items of bradykinesia) from the videos taken by Rater 1 and for the same 7 items from the corresponding videos taken by Rater 2. Sum K7-UPDRS scores and scores for individual items were compared between assessors in subsequent analysis.

Comparison of clinician and KELVIN scores

The mean of the 2 Sum K7-UPDRS scores derived from videos captured by Rater 1 and Rater 2 were compared to the mean of the 2 abbreviated Clinician scores C7-UPDRS of Rater 1 and Rater 2 for the equivalent 7 items.

Statistical analysis

Inter-rater reliability of the sum C18-UPDRS scores, inter-rater reliability of the sum K7-UPDRS scores and inter-rater reliability of the mean K7-UPDRS and mean C7-UPDRS scores were assessed using an intraclass correlation coefficient [4] method using a two-way, single measure, absolute-agreement random-effects model [21].

For intra-rater reliability of sum C18-UPDRS scores, an ICC method using a two-way, single measure absolute-agreement mixed-effects model was used [21, 22]. 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. Ninety-five percent confidence bounds were also computed for the ICCs using standard methods.

To minimize the impact of missing data, items that were missing from the clinical or KELVIN scores were imputed based on the mean scores for that participant's assessment for the non-missing data. For participants with more than 25% items missing, scores were excluded from the analysis. Reliability for individual MDS-UPDRS part 3 items was assessed using weighted kappa statistics [4].

RESULTS

Participant demographic data

The 22 participants in this study had a mean age of 69, (range 40–71), and a mean duration of disease of 11 years; 12 were male.

Inter and intra-rater variability

Inter-rater variability

Data are presented in Table 2. The inter-rater reliability for the sum C18-UPDRS scores was excellent in both the OFF (0.99, 95% CI 0.98–1.0) and ON (0.94 95% CI 0.78–0.98) medication conditions across raters.

Intra-rater reliability

The ratings by Rater 1 were repeated after a mean interval of 43 days (+-11 days). Intra-rater reliability for the sum C18-UPDRS scores was 0.98 (95% CI 0.94–0.99) and 0.92 (95% CI 0.82–0.97) for the OFF and ON scores.

C7-UPDRS scores inter rater reliability

The ICCs for inter-rater reliability and associated 95% confidence interval for the sum Subsection C7-UPDRS scores was excellent (0.97 (95% CI 0.93–0.99)) for the OFF scores and good (0.79 (95% CI 0.54–0.91)) for the ON scores.

K7-UPDRS scores inter rater reliability

Inter-rater reliability for the sum subsection K-UPDRS scores was good (0.80 (95% CI 0.57–0.91)) for the OFF scores and moderate (0.73 (95% CI 0.44–0.89)) for the ON scores.

Sum C7-UPDRS and K7-UPDRS score correlation

The ICCs for inter-rater reliability for the sum Subsection C7-UPDRS and K7-UPDRS scores was excellent (0.84 (95% CI 0.64–0.93)) for the OFF scores but poor (0.31 (95% CI -0.08–0.64)) for the ON scores.

Individual item inter-rater variability

Weighted Kappa statistics for the 7 individual MDS-UPDRS items are displayed in Table 3. Consistency between raters ranged from substantial-near perfect agreement for all of the specific C7-UPDRS items apart from Left finger taps (ON), right hand movements (ON), left toe taps (ON), and arise from chair (ON) which showed moderate agreement between raters.

Consistency between K-UPDRS scores for individual items was moderate for Finger taps & Hand movements and Arising from Chair in the OFF state. Agreement was lower for pronation /supination, Leg agility, Toe taps and Gait in the OFF state. In the ON state, there was moderate agreement for finger taps, and leg agility but very low agreement for Pronation/supination, arising from a chair and Gait. There were more missing data for Gait and Arising from a chair than the limb bradykinesia assessments.

Intra-rater reliability ranged from moderate to near perfect agreement for all of the 18 specific C-UPDRS items (see Table 4).

DISCUSSION

The aim of this work was to investigate whether KELVIN, an artificial intelligence platform, may be comparable to traditional clinical MDS-UPDRS scoring in terms of its consistency of ratings when assessing severity of PD. The main findings were that; KELVIN and clinician scores were extremely highly correlated for the OFF condition but poorly correlated for the ON condition. Nevertheless, KELVIN showed moderate-good inter-rater reliability for the sum of the subsection K7-UPDRS OFF and ON scores. The majority of K7-UPDRS scores for individual items showed moderate consistency, although some axial items, e.g., gait showed poor agreement between the 2 video assessments particularly in the ON condition.

These scores were in comparison to inter-rater reliability of the same subsection scores by clinicians, which was good-excellent for OFF and ON conditions, and with the majority of individual

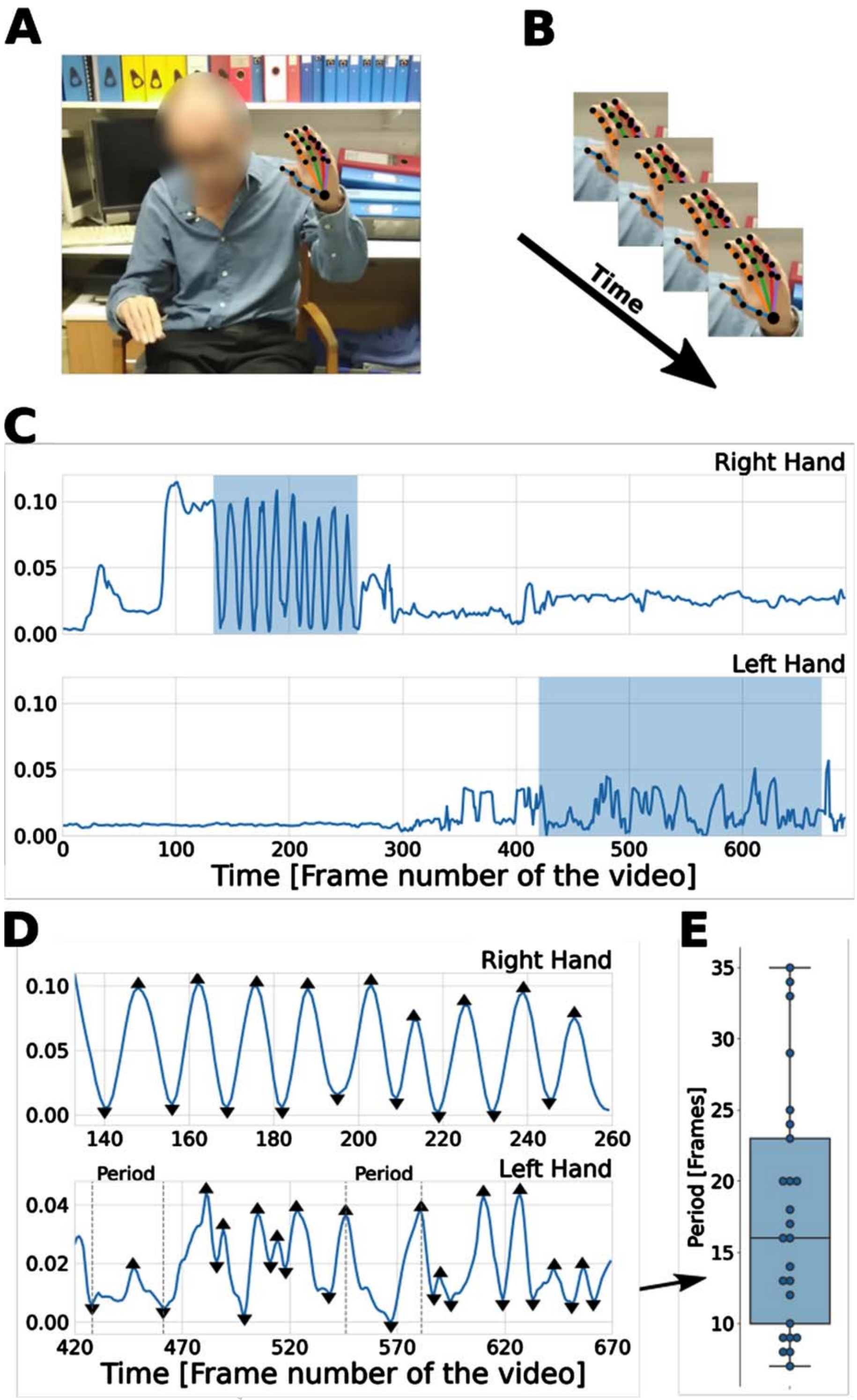

Fig. 1. (*Continued*)

Table 2
Inter- & intra rater reliability for total C18-UPDRS, and inter-rater reliability for C7-UPDRS & K7-UPDRS scores by Rater 1 and Rater 2

	Off medication	On Medication
C18-UPDRS scores- Inter rater reliability		
Mean (sd)	Rater 1 = 49.3 (15.6) Rater 2 = 48.6 (16.6)	Rater 1 = 15.9 (7.5) Rater 2 = 14.5 (6.7)
ICC [95% CI]	0.99 [0.98–1.0]	0.97 [0.88–0.98]
C18 UPDRS scores- Intra rater reliability		
Mean (sd)	OFF Time 1 = 49.0 (16.3) OFF Time 2 = 47.6 (15.2)	ON Time 1 = 16.2 (7.8) ON Time 2 = 16.5(6.2)
ICC [95% CI]	0.98 [0.94–0.99]	0.92 [0.82–0.97]
C7-UPDRS scores- Inter rater reliability*		
Mean (sd)	OFF Rater 1 = 21.1 (7.3) OFF Rater 2 = 21.2 (7.3)	ON Rater 1 = 7.1 (3.5) ON Rater 2 = 6.6 (3.4)
ICC [95% CI]	ICC:0.97 [0.93–0.99]	ICC: 0.79 [0.53–0.91]
Mean (Rater 1 & Rater 2) C7-UPDRS and Mean (Rater 1 & Rater 2) K7-UPDRS scores- Correlation coefficients		
Mean (sd)	OFF C7-UPDRS = 21.2 (7.3) OFF K7-UPDRS = 21.9 (8.1)	ON C7-UPDRS = 6.9 (3.3) ON K7-UPDRS = 10.3 (3.0)
ICC [95% CI]	0.84 [0.64–0.93]	0.31 [–0.08–0.64]
K7-UPDRS scores- Inter rater reliability		
Mean (sd)	OFF Rater 1 = 22.2 (9.4) OFF Rater 2 = 21.7 (7.6)	ON Rater 1 = 10.2 (4.5) ON Rater 2 = 10.4 (4.8)
ICC [95% CI]	0.80 [057–0.91]	0.73 [0.44–0.89]

C18-UPDRS, Clinician scored Movement Disorder Society Unified Parkinson's disease Rating Scale Part 3- total 18 items; C7-UPDRS, Clinician scored Movement Disorder Society Unified Parkinson's disease Rating Scale Part 3- restricted to 7 items; K7-UPDRS, Kelvin calculated Movement Disorder Society Unified Parkinson's disease Rating Scale Part 3- restricted to 7 items; CI, confidence intervals upper and lower bounds; ICC, intraclass correlation coefficient.
*Only includes patients with complete K7UPDRS scores for ease of comparison.

items showing excellent or near-perfect agreement, with some items of bradykinesia showing moderate agreement in the ON condition. Overall, intra-rater reliability showed excellent reliability for sum C18-UPDRS scores and moderate-near perfect agreement across all items. Taken together, these data indicate that KELVIN, in its current format, can provide extremely similar scores to the C7-UPDRS scores in the OFF-medication state, but produces consistent but different scores from the clinician ratings in the ON-medication state.

The findings in the present study for intra-rater reliability for C18-UPDRS scores are in line with previous research showing excellent intra-rater agreement [5, 7, 8]. However, inter-rater reliability for C18-UPDRS scores somewhat outperformed previous research, with scores in the present study showing good-excellent inter-rater agreement (0.88–0.98) compared to previous studies showing moderate-excellent reliability (0.65–0.91) [5–7]. Post et al. (2005) demonstrated lesser consistency between raters, when ratings of more senior movement disorder specialists were compared to less experienced movement disorder specialists. In addition, the raters in Palmer et al. (2010), who showed inter-rater agreement of 0.65 ICC, were not highlighted as being movement disorder specialists but dementia specialists, which may have meant that they had less experience administering the MDS-UPDRS. Further, raters noted as having expertise in movement disorders also showed similar agreement with that of scores in the present study, with ICC over 0.90 [7].

Fig. 1. Methods overview. A) The deep learning library OpenPose [2] was used to extract 25 body and 21 hand key points from each frame of video. B) Coordinates of the key points across the frames were used to construct time-series signals. C) An example of finger-tapping signals (i.e., distance between index finger tip and thumb tip) for right (top) and left (bottom) hand. In this case the right hand received a low severity score of 1, while the left hand received a high severity score of 4. The highlighted regions depict the regions of interest (ROIs); i.e., when the action was performed. D) Detected peaks and troughs on the signals of the two ROIs for the right hand (top) and left hand (bottom). Features were constructed from these signals. For example, the time between peaks corresponds to the time between successive finger taps. E) The distribution of periods (distance in frames between consecutive peaks and troughs) extracted from the lower panel (left hand signal) displayed in (D). Courtesy of Machine Medicine Technologies Ltd.

Table 3
Inter-rater reliability for individual C7-UPDRS & K7-UPDRS item scores by Rater 1 and Rater 2

	Weighted kappa scores for inter-rater agreement for 7 individual items of the C7-UPDRS		*Weighted kappa scores for inter-rater agreement for the 7 individual items of the K-UPDRS*	
	Rater 1 vs. Rater 2 OFF (SE)	Rater 1 vs. Rater2 ON (SE)	Rater 1 vs. Rater2 OFF (SE)	Rater 1 vs. Rater2 ON (SE)
Finger taps				
Right	N = 22 0.79 (0.08)	N = 22 0.75 (0.11)	N = 21 0.64 (0.12)	N = 19 0.52 (0.15)
Left	N = 22 0.80 (0.10)	N = 22 0.44 (0.17)	N = 22 0.63 (0.11)	N = 19 0.64 (0.13)
Hand movements				
Right	N = 22 0.75 (0.11)	N = 22 0.45 (0.19)	N = 21 0.63 (0.13)	N = 21 0.15 (0.16)
Left	N = 22 0.58 (0.14)	N = 22 0.67 (0.15)	N = 21 0.55 (0.11)	N = 21 0.46 (0.16)
Pronation/Supination				
Right	N = 22 0.66 (0.12)	N = 22 0.83 (0.10)	N = 20 0.11 (0.15)	N = 21 0.21 (0.17)
Left	N = 22 0.86 (0.08)	N = 22 0.79 (0.10)	N = 20 0.19 (0.13)	N = 21 –0.11 (0.13)
Toe Tap				
Right	N = 22 0.83 (0.07)	N = 22 0.78 (0.12)	N = 20 0.21 (0.15)	N = 19 0.25 (0.17)
Left	N = 22 0.73 (0.10)	N = 22 0.48 (0.13)	N = 20 0.47 (0.14)	N = 19 0.59 (0.11)
Leg Agility				
Right	N = 22 0.85 (0.07)	N = 22 0.67 (0.17)	N = 18 0.49 (0.14)	N = 20 0.63 (0.15)
Left	N = 22 0.86 (0.08)	N = 22 0.90 (0.10)	N = 19 0.46 (0.16)	N = 20 0.35 (0.17)
Arise from chair	N = 22 0.91 (0.05)	N = 22 0.55 (0.22)	N = 16 0.84 (0.06)	N = 14 0.10 (0.24)
Gait	N = 22 0.87 (0.09)	N = 22 0.86 (0.10)	N = 12 0.38 (0.18)	N = 14 0.08 (0.13)

The findings presented here may reflect the single center nature of the project, possible additional effort made by the raters to optimize their scores, and the very high levels of experience of all of the raters. Multi-center studies would be useful in determining whether KELVIN can reduce inconsistency of MDS-UPDRS scoring in comparison to less experienced clinicians across multiple centers, perhaps when under real-life time pressure to collect data and thus build on the current context-question from the present research.

In terms of individual items, the findings in the present study for intra-rater reliability show similar scores to previous research. The intra-rater variability for tremor and items of bradykinesia ranging from 0.53–0.84, are in line with previous studies showing agreement between 0.43–0.93 for tremor and bradykinesia [5, 7, 8]. Likewise agreement for rigidity items, posture, gait, facial expression, and speech showed a similar range of scores (0.63–0.82) to these items in previous measures of intra-rater reliability [5, 7, 8]. Thus the present study agrees with previous findings of excellent MDS-UPDRS part 3 intra-rater reliability.

The findings in the present study indicate that KELVIN does not yet improve upon an experienced clinician's scores of PD severity, particularly for gait. The low agreement in KELVIN scores for gait (both OFF/ON) contrasts a previous study that show KELVIN's ability to accurately assess gait in PD [19]. This may be due to the precision and complexity required when scoring gait on video. In a recent study, where raters scored the MDS-UPDRS via iPad tablets, this challenge has been highlighted as a weakness of video-scoring the MDS-UPDRS [23]. Improvement in Kelvin's algorithms with increased data exposure may improve the precision of its estimates. Another reason for the variability in gait scores

Table 4
Intra-rater reliability for individual C18-UPDRS item scores by Rater 1

	Rater 1 Time 1 vs. Time 2 Weighted kappa for individual items (SE)
Speech	0.67 (0.12)
Facial expression	0.82 (0.06)
Rigidity	
Neck	0.86 (0.05)
RUE	0.88 (0.05)
LUE	0.62 (0.09)
RLE	0.90 (0.04)
LLE	0.89 (0.07)
Finger taps	
Right	0.73 (0.07)
Left	0.66 (0.09)
Hand movements	
Right	0.76 (0.08)
Left	0.77 (0.08)
Pronation/Supination	
Right	0.78 (0.07)
Left	0.84 (0.06)
Toe Tap	
Right	0.74 (0.07)
Left	0.78 (0.07)
Leg Agility	
Right	0.64 (0.09)
Left	0.54 (0.12)
Arise from chair	0.88 (0.05)
Gait	0.63 (0.08)
FOG	0.94 (0.05)
Postural instability	0.90 (0.05)
Posture	0.70 (0.08)
Body Bradykinesia	0.78 (0.07)
Postural Tremor	
Right	0.76 (0.07)
Left	0.53 (0.09)
Action Tremor	
Right	0.65 (0.09)
Left	0.62 (0.13)
Resting Tremor	
RUE	0.84 (0.06)
LUE	0.75 (0.10)
RLE	0.67 (0.17)
LLE	0.77 (0.16)
Face/Neck	0.70 (0.10)
CON	0.79 (0.06)

RUE, right upper extremity; LUE, left upper extremity; RLE, right lower extremity; LLE, left lower extremity; CON, constancy of resting tremor; SE, standard error.

may have been the angle in which the camera was positioned to record the movement. Whilst positioned to the best of the raters' ability, the clinic where patients performed the gait assessments has a narrow corridor, and the impact of camera angle and distance from patient to camera may have had a more profound effect on automated gait assessment and rising from a chair scores resulting in greater numbers of missing data and greater variability in KELVIN gait scores (Number of missing datapoints quantified in Table 3). Greater experience of KELVIN usage should improve choice of camera and patient positioning to help mitigate this in the future.

The present study findings demonstrate that K7-UPDRS scores agree with clinician C7-UPDRS scores to an excellent degree in the OFF condition, and this shows a superior score to some studies of inter-rater reliability scores between clinicians [5, 6]. This shows that KELVIN was as accurate as clinicians at scoring the 7 items of the MDS-UPDRS when patients were OFF medication. However, KELVIN showed poor agreement with clinicians in the ON condition, which may be due to the majority of patients experiencing dyskinesia when ON medication. The impact of dyskinesia on movement may have interfered with KELVIN's scoring process and caused the system to assign higher scores to patients with dyskinesia. The higher mean of the average K7-UPDRS scores compared to the C7-UPDRS support this supposition and show that the raters scored patients 3.4 MDS-UPDRS points lower than the KELVIN platform in the ON condition. Whether clinician or KELVIN scores are a better measure of functional disability in the ON medication condition would be of interest. Further, this challenge of video assessment in the clinic setting may also translate to the home environment; with the added issue of possible lighting and space constraints and video assessment may also be more challenging for patients with severe motor symptoms to carry out without clinician support in the home setting. The FDA guidance provides a number of steps to validate a Digital Health Technology (DHT) on the population of interest to ensure that the DHT is fit-for-purpose for remote data collection use in a clinical investigation [24]. In line with the FDA guidance, KELVIN is able to consistently and appropriately measure a number of clinical symptoms of Parkinson's disease, which demonstrates that in the clinic, KELVIN is a fit-for-purpose DHT. Research conducted in the participant's homes could further validate the use of KELVIN in the home environment.

One limitation of KELVIN is that other aspects of C18-UPDRS are not yet usefully captured by K-UPDRS, therefore a full clinical picture (tremor, rigidity, speech, and postural instability) cannot yet be captured with KELVIN alone, which may reduce its competitiveness against conventional clinical assessment of PD severity, and it can in no way be used to replace expert clinical management. Platforms to assess speech and facial expression do also exist [25–27] and could be incorporated to provide a more complete automated clinical picture, which may reduce the potential biases of human raters.

Despite this, one relative strength of KELVIN is that it can measure multiple MDS-UPDRS items affecting different body parts as a single platform, which is something not yet possible with other AI PD severity measuring tools such as wearable sensors which can only measure one or two motor features of PD depending on where they are positioned [17, 28–30]. Attempts to measure all 18 items of the MDS-UPDRS part 3 may require the use of different wearable sensors for various items [28, 31], though these are often expensive and require high tech equipment which presents a disincentive for their routine use. A video-based AI tool as in the present study, presents itself as a more cost-effective and accessible approach than other types of AI technology, which may also allow for remote assessment and thus reduce the cost of attending appointments. Video-based assessment of PD severity may also prove to be useful in clinical trials, where items of the MDS-UPDRS may be carried out multiple times a year as an outcome measure, thus participants may not have to incur the costs and inconvenience of travel to attend appointments. This may be particularly useful for more disabled patients who have difficulty travelling or those who live far from the clinical trial testing site.

A major strength of KELVIN is its practical application. KELVIN incorporates intuitive software, with good usability, with all raters able to use the system with little guidance. This was useful in the present study's movement disorder clinic, and KELVIN supported clinician ratings by giving raters the ability to film patients undergoing MDS-UPDRS assessments and saving assessments in a single storage system. For example, anecdotal evidence from the present study demonstrated nurses having the ability to go back over items that were 'difficult to assess' and quality control check their scores. Furthermore, when considering the benefits of KELVIN for individual patient applications, the routine storage of videos in the KELVIN Cloud system can provide easy access and help clinicians review changes over time in individual patients.

Conclusions

In summary, the present research confirmed the practical application of KELVIN system to record, and store video recordings of the MDS-UPDRS part 3 and successfully analyze the bradykinesia, rising and gait items. Variability in scores still occur potentially depending on the human contributions such as camera position, angle as well as the different cameras, hardware and software included which may influence the consistency of the scores produced. When compared to conventional clinical scores captured by experienced raters all trained within a single center, the AI platform does not show a clear advantage.

Nevertheless, these data suggest that KELVIN shows promise, indeed it showed good agreement in overall K7-UPDRS scores between different videos, particularly in upper limb bradykinesia. The usefulness of the conventional MDS-UPDRS part 3 has depended on careful instructions being developed to accompany its use as well as teaching videos to ensure consistency of its application. Further iterations of the Kelvin platform, additional learning through greater amount of data captured, as well as simple but clear instructions regarding the standard approach to capture the videos should allow this tool to improve to the level required to complement and potentially improve upon the conventional clinical assessment of PD motor severity.

ACKNOWLEDGMENTS

We are grateful to all patients who participated in this study.

CONFLICT OF INTEREST

This study was funded by Innovate UK, Capital Enterprise, in a grant jointly awarded to UCL and Machine Medicine Technologies. Data were collected at the National Hospital for Neurology and Neurosurgery UCL Queen Square Institute of Neurology.

TF has received grants from National Institute of Health Research, Michael J Fox Foundation, John Black Charitable Foundation, Cure Parkinson's Trust, Innovate UK, Van Andel Research Institute and Defeat MSA. He has served on Advisory Boards for Voyager Therapeutics, Handl therapeutics, Living Cell Technologies, Bial, Profie Pharma and Peptron. He has received honoraria for talks sponsored by Bial, Profile Pharma, Boston Scientific and Novo Nordisk.

REFERENCES

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

[2] Vijiaratnam N, Simuni T, Bandmann O, Morris HR, Foltynie T (2021) Progress towards therapies for disease modification in Parkinson's disease. *Lancet Neurol* **20**, 559-572.

[3] Goetz CG, Poewe W, Rascol O, Sampaio C, Stebbins GT, Counsell C, Giladi N, Holloway RG, Moore CG, Wenning GK, Yahr MD, Seidl L (2004) Movement Disorder Society Task Force report on the Hoehn and Yahr staging scale: Status and recommendations. *Mov Disord* **19**, 1020-1028.

[4] Cicchetti DV, Allison T (1971) A new procedure for assessing reliability of scoring EEG sleep recordings. *Am J Electroneurodiagnostic Technol* **11**, 101-110.

[5] Post B, Merkus MP, de Bie RM, de Haan RJ, Speelman JD (2005) Unified Parkinson's disease rating scale motor examination: Are ratings of nurses, residents in neurology, and movement disorders specialists interchangeable? *Mov Disord* **20**, 1577-1584.

[6] Palmer JL, Coats MA, Roe CM, Hanko SM, Xiong C, Morris JC (2010) Unified Parkinson's Disease Rating Scale-Motor Exam: Inter-rater reliability of advanced practice nurse and neurologist assessments. *J Adv Nurs* **66**, 1382-1387.

[7] Bennett DA, Shannon KM, Beckett LA, Goetz CG, Wilson RS (1997) Metric properties of nurses' ratings of parkinsonian signs with a modified Unified Parkinson's Disease Rating Scale. *Neurology* **49**, 1580-1587.

[8] Siderowf A, McDermott M, Kieburtz K, Blindauer K, Plumb S, Shoulson I (2002) Test-retest reliability of the Unified Parkinson's Disease Rating Scale in patients with early Parkinson's disease: Results from a multicenter clinical trial. *Mov Disord* **17**, 758-763.

[9] Heldman DA, Giuffrida JP, Chen R, Payne M, Mazzella F, Duker AP, Sahay A, Kim SJ, Revilla FJ, Espay AJ (2011) The modified bradykinesia rating scale for Parkinson's disease: Reliability and comparison with kinematic measures. *Mov Disord* **26**, 1859-1863.

[10] Richards M, Marder K, Cote L, Mayeux R (1994) Interrater reliability of the Unified Parkinson's Disease Rating Scale motor examination. *Mov Disord* **9**, 89-91.

[11] Williams S, Zhao, Z, Hafeez A, Wong DC, Relton SD, Fang H, Alty JE (2020). The discerning eye of computer vision: Can it measure Parkinson's finger tap bradykinesia? *J Neurol Sci* **416**, 117003.

[12] Monje MHG, Dominguez S, Vera-Olmos J, Antonini A, Mestre TA, Malpica N, Sanchez-Ferro A (2021) Remote evaluation of Parkinson's disease using a conventional webcam and artificial intelligence. *Front Neurol* **12**, 742654.

[13] Butt AH, Rovini E, Fujita H, Maremmani C, Cavallo F (2020) Data-driven models for objective grading improvement of Parkinson's disease. *Ann Biomed Eng* **48**, 2976-2987.

[14] Dai H, Cai G, Lin Z, Wang Z, Ye Q (2021) Validation of inertial sensing-based wearable device for tremor and bradykinesia quantification. *IEEE J Biomed Health Inform* **25**, 997-1005.

[15] Iakovakis D, Chaudhuri KR, Klingelhoefer L, Bostantjopoulou S, Katsarou Z, Trivedi D, Reichmann H, Hadjidimitriou S, Charisis V, Hadjileontiadis LJ (2020) Screening of Parkinsonian subtle fine-motor impairment from touchscreen typing via deep learning. *Sci Rep* **10**, 12623.

[16] Kleinholdermann U, Wullstein M, Pedrosa D (2021) Prediction of motor Unified Parkinson's Disease Rating Scale scores in patients with Parkinson's disease using surface electromyography. *Clin Neurophysiol* **132**, 1708-1713.

[17] Channa A, Ifrim RC, Popescu D, Popescu N (2021) A-WEAR bracelet for detection of hand tremor and bradykinesia in Parkinson's patients. *Sensors (Basel)* **21**, 981.

[18] Goetz CG, Stebbins GT, Luo S (2020) Movement Disorder Society-Unified Parkinson's disease rating scale use in the Covid-19 era. *Mov Disord* **35**, 911.

[19] Rupprechter S, Morinan G, Peng Y, Foltynie T, Sibley K, Weil RS, Leyland LA, Baig F, Morgante F, Gilron R, Wilt R, Starr P, Hauser RA, O'Keeffe J (2021) A clinically interpretable computer-vision based method for quantifying gait in Parkinson's disease. *Sensors (Basel)* **21**, 5437.

[20] Morinan G, Peng Y, Rupprechter S, Weil RS, Leyland LA, Foltynie T, Sibley K, Baig F, Morgante F, Gilron Re, Wilt R, Starr P, O'Keeffe J (2022) Computer-vision based method for quantifying rising from chair in Parkinson's disease patients. *Intell Based Med* **6**, 100046.

[21] Koo TK, Li MY (2016) A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *J Chiropr Med* **15**, 155-163.

[22] Shrout PE, Fleiss JL (1979) Intraclass correlations: Uses in assessing rater reliability. *Psychol Bull* **86**, 420-428.

[23] Sekimoto S, Oyama G, Hatano T, Sasaki F, Nakamura R, Jo T, Shimo Y, Hattori N (2019) A randomized crossover pilot study of telemedicine delivered via iPads in Parkinson's disease. *Parkinsons Dis* **2019**, 9403295.

[24] Food & Drug Administration Oncology Center of Excellence (2021) Digital Health Technologies for Remote Data Acquisition in Clinical Investigations. Downloaded from; https://www.fda.gov/regulatory-information/search-fda-guidance-documents/digital-health-technologies-remote-data-acquisition-clinical-investigations. Accessed August 2022.

[25] Carrón J, Campos-Roca Y, Madruga M, Pérez CJ (2021) A mobile-assisted voice condition analysis system for Parkinson's disease: Assessment of usability conditions. *Biomed Eng Online* **20**, 114.

[26] Dastjerd NK, Sert OC, Ozyer T, Alhajj R (2019) Fuzzy classification methods based diagnosis of Parkinson's disease from speech test cases. *Curr Aging Sci* **12**, 100-120.

[27] Jin B, Qu Y, Zhang L, Gao Z (2020) Diagnosing Parkinson disease through facial expression recognition: Video analysis. *J Med Internet Res* **22**, e18697.

[28] Adams JL, Dinesh K, Snyder CW, Xiong M, Tarolli CG, Sharma S, Dorsey ER, Sharma G (2021) A real-world study of wearable sensors in Parkinson's disease. *NPJ Parkinsons Dis* **7**, 106.

[29] Brognara L, Palumbo P, Grimm B, Palmerini L (2019) Assessing gait in Parkinson's disease using wearable motion sensors: A systematic review. *Diseases* **7**, 18.

[30] Pardoel S, Kofman J, Nantel J, Lemaire ED (2019) Wearable-sensor-based detection and prediction of freezing of gait in Parkinson's disease: A review. *Sensors (Basel)* **19**, 5141.

[31] Zhang H, Li C, Liu W, Wang J, Zhou J, Wang S (2020) A multi-sensor wearable system for the quantitative assessment of Parkinson's disease. *Sensors (Basel)* **20**, 6146.