CORA Cork Open Research Archive
Cartlann Taighde Oscailte Chorcaí

Title	Inter-rater reliability of hand motor function assessment in Parkinson's disease: impact of clinician training
Authors	Kenny, Lorna;Azizi, Zahra;Moore, Kevin;Alcock, Megan;Heywood, Sarah;Jonsson, Agnes;McGrath, Keith;Foley, Mary J.;Sweeney, Brian;O'Sullivan, Sean S.;Barton, John;Tedesco, Salvatore;Sica, Marco;Crowe, Colum;Timmons, Suzanne
Publication date	2024
Original Citation	Kenny, L., Azizi, Z., Moore, K., Alcock, M., Heywood, S., Jonsson, A., McGrath, K., Foley, M.J., Sweeney, B., O'Sullivan, S., Barton, J., Tedesco, S., Sica, M., Crowe, C. and Timmons, S. (2024) 'Inter-rater reliability of hand motor function assessment in Parkinson's disease: Impact of clinician training', Clinical Parkinsonism & Related Disorders, 11, 100278 (7pp). https://doi.org/10.1016/j.prdoa.2024.100278
Type of publication	Article (peer-reviewed)
Link to publisher's version	10.1016/j.prdoa.2024.100278
Rights	© 2024, the Authors. Published by Elsevier Ltd. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/) - https://creativecommons.org/licenses/by/4.0/
Download date	2026-11-08 09:10:28
Item downloaded from	https://hdl.handle.net/10468/16799

UCC
University College Cork, Ireland
Coláiste na hOllscoile Corcaigh

PRDOA 11 (2024) 100278

Contents lists available at ScienceDirect

Clinical Parkinsonism & Related Disorders

journal homepage: www.sciencedirect.com/journal/clinical-parkinsonism-and-related-disorders

Original Articles

Inter-rater reliability of hand motor function assessment in Parkinson's disease: Impact of clinician training

Lorna Kenny [a,*], Zahra Azizi [a], Kevin Moore [a], Megan Alcock [b], Sarah Heywood [b], Agnes Jonsson [b], Keith McGrath [b], Mary J. Foley [c], Brian Sweeney [d], Sean O'Sullivan [d], John Barton [e], Salvatore Tedesco [e], Marco Sica [e], Colum Crowe [e], Suzanne Timmons [a]

[a] Centre for Gerontology and Rehabilitation, School of Medicine, University College Cork, Cork T12 XH60, Ireland
[b] Mercy University Hospital, Cork T12 WE28, Ireland
[c] Cork Stroke Support Centre, Cork T12 AKA4, Ireland
[d] Neurology Department, Bon Secours Hospital, Cork T12 DV56, Ireland
[e] Tyndall National Institute, University College Cork, Lee Maltings, Prospect Row, Cork T12 AV22, Ireland

ARTICLE INFO

Keywords:
Parkinson's disease
Assessment
Inter-rater reliability
Variability
MDS-UPDRS

ABSTRACT

Medication adjustments in Parkinson's disease (PD) are driven by patient subjective report and clinicians' rating of motor feature severity (such as bradykinesia and tremor).

Objective: As patients may be seen by different clinicians at different visits, this study aims to determine the inter-rater reliability of upper limb motor function assessment among clinicians treating people with PD (PwPD).

Methods: PwPD performed six standardised hand movements from the Movement Disorder Society's Unified Parkinson's Disease Rating Scale (MDS-UPDRS), while two cameras simultaneously recorded. Eight clinicians independently rated tremor and bradykinesia severity using a visual analogue scale. We compared intraclass correlation coefficient (ICC) before and after a training/calibration session where high-variance participant videos were reviewed and MDS-UPDRS instructions discussed.

Results: In the first round, poor agreement was observed for most hand movements, with best agreement for resting tremor (ICC 0.66 bilaterally; right hand 95 % CI 0.50–0.82; left hand: 0.50–0.81). Postural tremor (left hand) had poor agreement (ICC 0.14; 95 % CI 0.04–0.33), as did wrist pronation-supination (right hand ICC 0.34; 95 % CI 0.19–0.56). In post-training rating exercises, agreements improved, especially for the right hand. Best agreement was observed for hand open-close ratings in the left hand (ICC 0.82, 95 % CI 0.64–0.94) and resting tremor in the right hand (ICC 0.92, 95 % CI 0.83–0.98). Discrimination between right and left hand features by raters also improved, except in resting tremor (disimprovement) and wrist pronation-supination (no change).

Conclusions: Clinicians vary in rating video-recorded PD upper limb motor features, especially bradykinesia, but this can be improved somewhat with training.

1. Introduction

Parkinson's disease (PD) is a neurodegenerative disorder that causes progressive decline in motor function and non-motor symptoms [1], affecting over 8.5 million people worldwide [2].

The Movement Disorder Society's Unified PD Rating Scale (MDS-UPDRS) is the gold-standard clinical assessment tool for PD, providing instructions for evaluating hand function (part III of the scale), where patients perform standardised movements, and clinicians assign severity scores based on the degree of motor function impairment [3]. Reliability has been demonstrated in previous studies (e.g. Goetz et al., 2008) [4], including external validation [5], nonetheless, the subjective nature of motor function rating by clinicians introduces potential variations in interpretation, even when strictly adhering to the guidelines [6]. Utilising the scale correctly depends on the skills of clinicians, where variations in ratings can stem from individual biases and differing assessments of specific motor functions. As patients undergo evaluations by different clinicians at different visits, these divergent scores can

* Corresponding author at: Centre for Gerontology and Rehabilitation, School of Medicine, University College Cork, St Finbarr's Hospital, Douglas Road, Cork T12 XH60, Ireland.
E-mail address: lorna.kenny@ucc.ie (L. Kenny).

https://doi.org/10.1016/j.prdoa.2024.100278
Received 23 July 2024; Received in revised form 9 October 2024; Accepted 20 October 2024
Available online 28 October 2024
2590-1125/© 2024 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

influence the evaluation of disease progression or treatment response [6].

Inter-rater reliability (IRR) refers to the reliability of data collected by multiple evaluators who assess the same individuals during a single instance [7]. It helps determine whether a measurement instrument yields accurate results that clinicians can rely on to make decisions. While offering an impartial evaluation of motor severity, the IRR of the MDS-UPDRS poses possible issues, especially in multicentre trials for PD and in long-term studies involving multiple raters assessing the same patient at different intervals [4]. It may not be adequately reliable for tracking a patient over time due to psychometric concerns such as; measurement of motor symptoms and impact in early PD, and misalignment in MDS-UPDRS-III items implies a lack of sensitivity in detecting variations and clinical change [8]. The MDS-UPDRS-III change scores display considerable variability due to error (within-subject reliability ranged from 0.13 to 0.62), highlighting an urgency for more dependable assessment [9].

Accurate assessment of people with PD (PwPD) is vital for providing optimal care support and treatment [10], monitoring disease progression [11], and advancing research [10].

A number of technologies have been used for supporting clinicians with the assessment of motor symptoms in PwPD. These technologies vary from very low-cost and unobtrusive systems to expensive technology requiring the set-up of specific labs [12]. Examples could include digitalised tools (such as smart pens) for handwriting assessment [13], smartphones (equipped with inertial sensors, Global Positioning Systems, audio, and camera) for real-world assessment in free-living settings [14], force plates [15], and pressure mats/insoles for evaluation of the quality of gait [16], cameras for monitoring patients' movements remotely at home or in controlled lab settings [17], motion capture systems (such as Vicon, or Microsoft Kinect) [18], physiological sensors (such as electrocardiogram, electromyography, electroencephalogram) [19], often combined within wearable sensors (generally equipped with inertial sensors) for monitoring remotely body movements, gait, and symptoms [20].

Going forward, PD care is likely to involve wearable devices [21], due to their low-cost, portability, and unobtrusiveness, and achieving accurate calibration depends on precise input from expert clinicians [22]. An online program offers teaching videos of the MDS-UPDRS and has shown to be an asset in improving both inter-rater and longitudinal reliability [23]. However, the effectiveness of learning may be impacted as videos are often watched in isolation and do not provide direct instructions or support. Interactive and team-based learning environments likely offer a more comprehensive experience and may foster deeper understandings [24]. Having direction and opportunity to ask questions in real time may be helpful for understanding and putting complex assessments like the MDS-UPDRS into practice [25,26]. To the best of the authors knowledge, no studies have investigated the potential improvement in IRR through a group calibration exercise.

This study aimed to determine the IRR of hand motor function assessment among clinicians who treat individuals with PD at specialised clinics and whether a group calibration exercise could enhance IRR. This research is part of a broader European-funded initiative called SENDoc (Smart sENsor Devices fOr rehabilitation and Connected health). (https://sendoc.interreg-npa.eu/). The primary objective was to assess the precision of wearable devices in measuring hand motor functions in PwPD, with the goal of developing a novel device that captures PD-related health information. Part of assessing devices is understanding the accuracy of examination by experts, alongside current practice, and to what degree this can be improved.

2. Methods

2.1. Participants

PwPD were recruited through branches of the Parkinson's Ireland and a specialised PD clinic in Munster, Ireland. Those interested received an information sheet about the study and researchers' contact details. Prospective participants contacted the team, allowing for detailed explanation of the study and the opportunity to ask questions. Inclusion criteria included individuals aged 50 and above with a confirmed PD diagnosis. A criterion-based theoretical sampling strategy ensured representation across different age categories: 50–60, 61–70, 71–80, and 80+, and a mix of male and female participants. Additionally, two spouses of a PwPD were invited to mimic a scenario where some patients have no abnormal upper limb motor status. Data collection occurred at Tyndall National Institute and St Finbarr's Hospital in Cork, Ireland. Participants provided informed consent and access to their personal data. The study received ethical approval from the Clinical Research Ethics Committee at University College Cork, number ECM 4 (a) 16/10/19.

2.2. Data collection

Demographic and clinical information collected included participant details such as age, gender, height, weight, the currently affected side (right, left, or both), and PD stage determined by the Hoehn and Yahr. The number of years since diagnosis and the participant's current subjective ON/OFF state were recorded.

In the assessment phase of the research, participants were seated in a desk chair in the study site. Researchers attached electromyography electrodes to each palm, and on the back of each forearm. Inertial measurement units were attached to the back of each hand. Following highly standardised instructions (Table 1 in the Supplementary File), participants executed six hand movements from the motor section of the MDS-UPDRS-III. Tasks included; resting tremor for 30 s, postural tremor for 15 s, and kinetic tremor exercises, five times per hand. Additionally, participants performed 30 rapid finger taps per hand, 15 hand opening and closing movements per hand, and 10 wrist pronation and supination movements per hand, all with arms extended in front of their body.

Fig. 1 in the Supplementary File shows a participant performing a hand movement assessment. Movements were recorded simultaneously from a frontal and 45-degree angle perspectives (to ensure one arm did not obstruct the other) using two cameras.

To ensure anonymity, facial blurring was applied to videos. The six-hand movements were then presented to the raters, as a short video clip, in random order to minimise potential bias.

2.3. Raters and severity scoring

Raters with clinical experience of PD were recruited through regional networks known to the research team, and allowed for representation of individuals with diverse experiences. Eight clinicians, consisting of two geriatricians (who ran a specialised PD clinic), two neurologists (one with a special interest in PD; one highly experienced in neurology), one PD specialist nurse, and three senior geriatric registrars (with at least 3 months' experience at a PD clinic), were initially recruited for the first round of data collection.

Raters independently reviewed the videos in accordance with the guidelines outlined by the MDS-UPDRS, and provided individual assessments for the severity of tremor and bradykinesia in each participant. Severity was marked in pen on a visual analogue scale (VAS) to generate continuous rather than categorical data. Some parts of the MDS-UPDRS, like rigidity and tremor amplitude, have clear criteria, making it easier for raters to agree. However, items related to bradykinesia, where terms like "slight", "mild", and "moderate" can lead to different interpretations among raters, and subsequently is more prone to variability [27]. The VAS was presented as a 100 mm black horizontal line on paper, with "Normal" printed to the left and "Severe" printed to the right, with no internal markings or numbers. Raters were instructed to draw a vertical line on the VAS to indicate severity ratings for each specific hand movement, separately for each hand (right and left) and

for each motor task. Researchers measured the distance from the left edge of the VAS line to the rater-marked vertical line of severity, with measurements recorded to the nearest millimetre. Ratings placed at the far left or right of the line were recorded as 0 or 100, respectively.

Following analysis of the initial ratings, clinicians convened in a 2-hour "calibration" session to explore variance between raters. As a group, four participant videos exhibiting high variance were reviewed and the MDS-UPDRS instructions were discussed. Subsequently, 3–5 weeks later, six clinicians independently rated a second set of videos featuring 10 new participants. Two clinicians (one neurologist and one senior geriatric registrar) were unable to complete due to work unforeseen commitments.

2.4. Statistical analysis

The primary evaluation metric was the intraclass correlation coefficient (ICC), a widely adopted measure for determining IRR, and is utilised in studies focusing on IRR and agreement analysis [28]. ICC was calculated for each movement in every rating round, as an estimation of the level of agreement among raters, and was calculated for each movement using the iccNA function in the "irrna" package (v0.2.3) in R (R Core Team (2022) https://www.R-project.org/) and 95 % confidence intervals were reported. A 2-way random effects model without interaction was utilised, and the results for absolute agreement [ICC(A,1)] were extracted. Koo and Li (2016) categorise ICC values as; below 0.50: poor reliability; 0.50 to 0.75: moderate reliability; 0.75–0.90: good reliability; and above 0.90: excellent reliability. Generally, for routine motor assessments (e.g., bradykinesia, tremor), an ICC above 0.75 is often deemed clinically acceptable for research or diagnostic purposes. For critical diagnostic decisions or clinical trials, a higher ICC of above 0.90 is preferred [29]. Lower ICC values may reflect not only poor agreement between raters but also experimental errors, rater biases, or lack of variability among subjects [29]. For PD assessments in real-world settings, high IRR between clinicians assessing the same motor tasks is essential for reliable monitoring of the disease, and improved clinical decision making [7]. Wilcoxon signed tests were used to compare continuous variables across two rounds. The level of significance was established as 95 % ($p < 0.05$).

Table 2
Participant demographics.

	1st Round N = 20	2nd Round N = 10
Gender		
Male	12	4
Female	8	6
Age		
Median	72	71
Range	55–81	60–80
Weight (kg)		
Median	73	71.5
Range	57–102	57–102
PD specific features	**N = 18**	**N = 10**
Subjective ON/OFF State		
ON	14	8
OFF	2	2
Does not experience/Not aware	2	0
Self-reported affected side		
Left	9	4
Right	6	6
Bilateral	2	0
Participant unsure	1	0
Hoehn and Yahr Stage		
1	0	2
2	11	5
3	5	3
4	2	0
Years since diagnosis		
Median	4	4.5
Range	1–14	1–14

3. Results

The first round of ratings included 18 PwPD and 2 healthy controls, ranging in age from 55 to 81, Hoehn and Yahr (H&Y) stage 2–4, with 61.1 % (11/18) at stage 2. The predominantly affected body side was left in nine participants (self-report), right in six and bilateral in two, while one was unsure. For fourteen participants, all assessments were performed in a self-reported clear ON state; in two participants, they felt clearly OFF throughout; the remainder (n = 2) did not experience clear ON/OFF periods in their opinion.

The second round of ratings included ten PwPD, ranging in age from 60 to 80, at H&Y stage 1–3, with the right side of the body more often affected (6:4). Table 2 contains participant summary statistics. No statistical difference existed between the two groups in any of the following: age (Wilcoxon signed rank test, W = 96.5, P = 0.89), H&Y stage (W = 104.5, P = 0.45), and years since diagnosis (W = 102, P = 0.58).

In round 1 (R1; n = 20 participants; n = 8 raters), mean severity scores (arbitrary units) from individual raters for assessments of bradykinesia-finger tapping, Hand Open-Close (HOC), and Wrist Pronation-Supination (WPS)-were higher than ratings for tremors, in

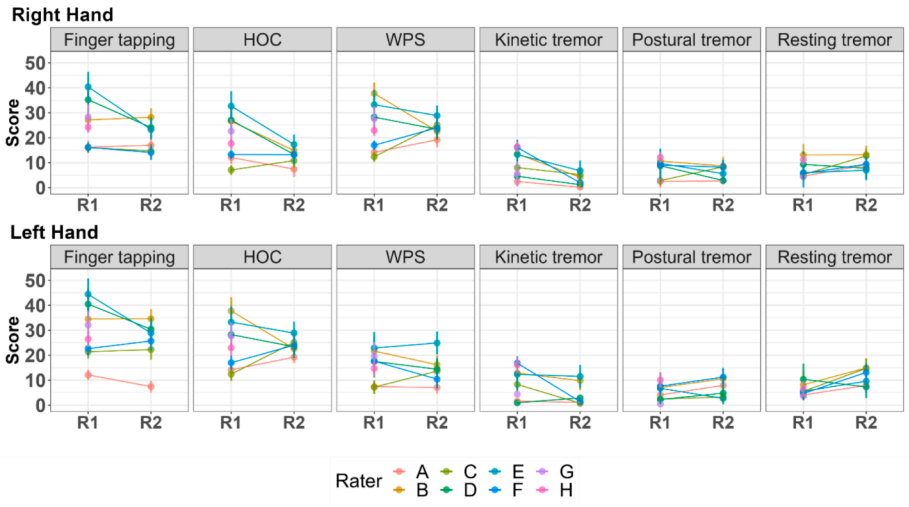


Fig. 2. Mean severity score (arbitrary units), with standard error, from each rater for each motor task in round 1 (R1) and round 2 (R2), presented as data for right hands (above) and for left hands (below).

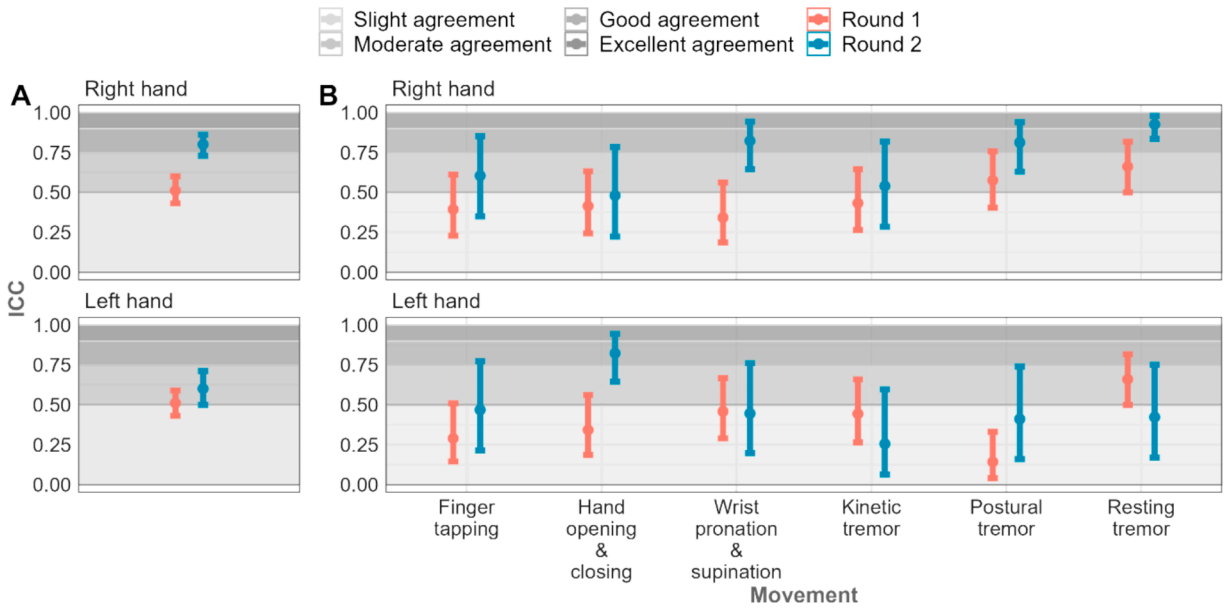


Fig. 3. ICCs between raters for round 1 and round 2 for (A) all data (B) each movement. In both figures, data for the right hand is presented above, and the left hand below.

both hands (right- and left-hand data presented separately; Fig. 2 in text and Table 3 in Supplementary File). In round 2 (R2; n = 10 participants; n = 6 raters), this difference is still present but less marked.

3.1. Agreement between raters

In round 1, there was moderate-to-poor agreement in ratings (ICC) among clinicians across the 240 discrete hand movement videos (overall ICC 0.51), with best agreement for resting tremors (right hand: ICC 0.66; 95 % CI 0.50–0.82; left hand: ICC 0.66; CI 0.50–0.81). Postural tremor in the left hand (ICC 0.14; 95 % CI 0.04–0.33) and WPS in the right hand (ICC 0.34; 95 % CI 0.19–0.56) had the least agreement (Fig. 3 in text and Table 4 in Supplementary File).

For post-training rating (round 2; n = 6 clinicians; 10 PwPD; 120 discrete hand movements) agreements improved generally (overall ICC 0.70), especially in the right hand (noting confidence intervals are wider due to smaller rater numbers; Fig. 3 in text; Table 4 in Supplementary Files). Best agreement was for HOC in the left hand (ICC 0.82; 0.64–0.94) and resting tremor in the right hand (ICC 0.92; 0.83–0.98). Least agreement was for kinetic tremor in the left hand (ICC 0.25; 0.06–0.60) and HOC in the right hand (ICC 0.48; 0.22–0.78).

3.2. Degree of relatedness in the ratings of hands

In round 1, there was a high degree of relatedness in the ratings of both hands for kinetic tremor severity (ICC 0.80; 95 % CI 0.74–0.85), and a moderate degree for finger tapping (ICC 0.74; 95 % CI 0.66–0.81) and HOC (ICC 0.71; 0.61–0.79). Therefore, raters appeared to judge both hands similarly (even though PD is by definition asymmetrical), and 15/18 participants identified a "worst" side. In round 2, this decreased significantly for three assessments. This may reflect participant characteristics (i.e., may have been more obvious asymmetry as not all identified a "worst" side) but suggests more independent rating by clinicians of each hand (Fig. 4 in text and Table 5 in Supplementary File).

4. Discussion

This study shows substantial variability in experienced clinicians' assessments of PD motor features while examining video-recorded hand

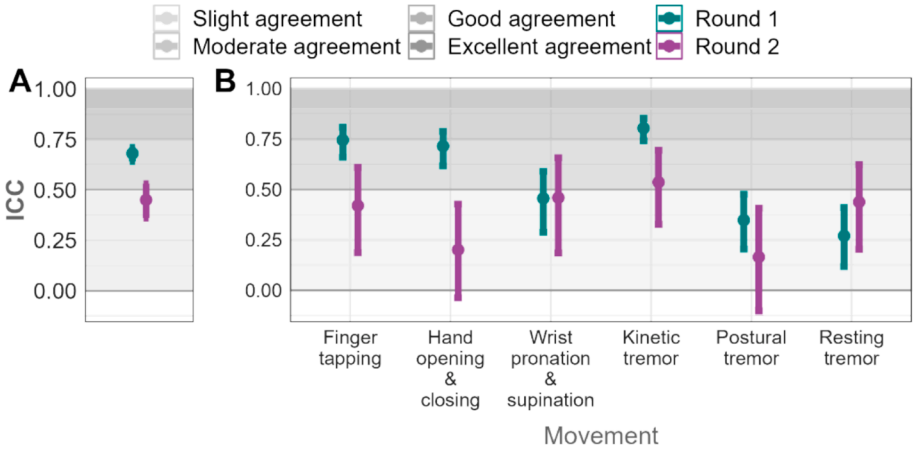


Fig. 4. Relatedness of the severity scoring between two hands: (A) for all the data, (B) for each movement. A higher ICC indicates a greater relatedness between scores for the left and right hands in a single participant.

movements, where overall agreement (all movements combined) between raters was 0.51 (sub-item's ICC ranged from 0.14 to 0.66). However, variability can be mitigated to an extent through a training and calibration session, resulting in overall ICC of 0.70, with a range from 0.25 to 0.92.

Although earlier studies show the ICC coefficients among MDS-UPDRS movements were high (ICC 0.90; [30])), IRR exhibited a variation among different studies. Richards et al. (1994) tested the IRR of the UPDRS (precursor to the MDS-UPDRS) [25] and reported a high ICC for resting tremor (0.84) but interestingly, the movement with best reliability was wrist pronation and supination which contrasts with our study where this movement displayed one of the least reliable assessments. Overall, their ICCs were higher than ours for similar movements. Meanwhile, there was only moderate agreement for finger tapping rating between neurologists (ICC 0.53; [31]). Also, Morinan et al. (2023) tested bradykinesia from the MDS-UPDRS encompassing various motor movement ratings [32].

Poorer agreement in our study may be attributed to several factors. Firstly, and importantly, previous studies used the ordinal 0–4 scale of the MDS-UPDRS, whereas we used continuous data on a VAS scale, which is arguably a more "pure" data measure where differences between raters are more apparent. Secondly, previous findings demonstrated a greater variation in rater scores when a limited number of movements were assessed in isolation, rather than alongside other clinical assessments [33], such as our study. Thirdly, in some of these studies [25,34], few raters are used (4 and 3, respectively), compared to eight in ours, although other studies had larger panels of raters (e.g. 15 in Morinan et al., 2023 [32]).

Following group calibration clinicians reported a more independent rating of each hand. Most PD is asymmetric (i.e., idiopathic PD) and even when bilateral, the initially/worst affected side remains most affected [35]. This lateralisation is diagnostically important in suggesting PD aetiology (i.e. typical; versus atypical or more likely drug-induced), and right side predominant motor features have a more rapid progression [36]. Meanwhile, patients with their dominant side more affected have significantly poorer postural control [37]. Thus, ability to lateralise motor features is prognostically important, and so ability to improve left–right motor feature discrimination through training is a bonus. Several studies using ordinal scales have reported variability in motor symptom assessments among clinicians. For instance, Ren et al. (2021) and Goetz et al. (2008) emphasize the challenges of achieving consistent ratings with traditional ordinal scales [4,38]. Our findings align with this research, extending it by demonstrating that VAS, which provide continuous and granular data, demonstrate significant inter-rater variability. VAS are commonly used clinically for assessing pain [39] and quality of life (e.g. EuroQol's EQ5D). Literature indicates that both ordinal and VAS outcomes are susceptible to subjective interpretation in assessing individuals with PD, across domain such as pain [40], global severity [41], and quality of life [42]. An advantage of VAS in research, and also clinically, is its ability to detect subtle symptom severity graduations and to avoid language that might cause rater bias (such as "severe"). While two assessors may both choose "moderate" for severity of a motor task (and hence be judged to rate the same), a VAS can reveal differences in rating if one scores 50 for example and another 70. In our study, following training, IRR significantly improved for movements such as hand open-close and resting tremor, mirroring findings by Goetz et al. (2008) that training enhances reliability for ordinal scales. Notably, our results revealed non-uniform effects in resting tremor ratings post-training, specifically, one hand may exhibit an agreement, while the other may show a decrease, diverging from previous studies that typically report uniform improvements [4]. Ultimately, while VAS offers continuous data sensitive to subtle changes, which is useful for research, ordinal scales may be easier to rate and compare in clinical practice.

The MDS-UPDRS is recognised globally in routine clinical practice and is the most commonly accepted reference standard for assessing PwPD, providing clear categorical data that is easy to analyse, familiar, and consistent. Despite already stated limitations in categorical scales, it is still the most preferred approach to assessing PD for many. Given this already established status, prioritising training in the proper use is especially justified to ensure clinicians use consistently [43]. Our findings support this, as after a group calibration exercise, generally the ICCs improved. The most difficult case to rate is the subject with least impairment and raters had the greatest difficulty with the mildest impairment, revealing that training is especially important in early PD [43]. Conversely, we found that the lower the mean scores for a movement, the higher the apparent ICC; for example, resting tremor. Reasons for this are not clear.

Enhancing IRR in PD assessments can support consistent identification of key symptoms like tremor, bradykinesia, and rigidity, leading to timely diagnoses, better interventions, and improved outcomes [10,11,26]. Reliable symptom assessment ensures accurate medication dose adjustments [44], and supports therapeutic decisions like deep-brain stimulation (DBS) candidacy, where a high IRR is crucial for assessing levodopa response, typically requiring a 33 % improvement in MDS-UPDRS scores [45]. Additionally, improving IRR enhances multidisciplinary care by aligning specialists on treatment needs. However, clinical assessments may miss the slight variations in movement that are indicative of early-stage PD, or overlook fluctuations in symptoms and real-life functioning. It is likely that objective, device-aided monitoring would be needed for such precision.

Technology and low-cost computer systems [12–20] offer an opportunity to improve IRR in PD care by reducing subjective variations. These systems can augment consistent assessments across clinical locations, and raters, minimising human error. They also offer feedback for clinician training and ensure reliable data collection, improving the consistency of care. While some of the technologies have shown promising results in terms of feasibility and accuracy, their selection by clinicians depends on the specifications required for specific clinical scenarios and end-users (e.g., remote monitoring, cost, patients' digital skills, assessment frequency).

5. Limitations

The study included a limited number of participants; a larger sample would provide greater understanding of IRR. The distribution of severity among participants was not Gaussian (normal), and the majority had mild to moderate severity PD (people with severe disease found it more challenging to participate), which may limit applicability to more advanced stages. Older people, minorities, and those with comorbidities are often underrepresented in research, limiting the ability to generalise study results [46]. Expanding the findings to a more diverse population could provide deeper insights into motor assessment variability in PD, as cultural and demographic factors affect disease presentation and access to care. There was a gender difference between the rounds, highlighting a potential variability that should be considered. Raters were mostly from one region which might limit the variety of experiences, and although there was a reasonable mix of expertise/discipline (three trainees, four consultants, one nurse), raters were more representative of geriatric medicine than neurology.

Two raters withdrew for the second round due to unforeseen professional commitments, and the second round had less participants (i.e., 8 raters rated 20 participants, and then 6 rated 10 participants). To ensure reliability, we compared the ICC from the same six raters in round 1 to the ICC from these six raters in round 2. Overall ICC for the six raters in round 1 was 0.46 (95 % CI [0.39–0.53]) and 0.70 (95 % CI [0.64–0.77]) in round 2. The trend of improvement from round 1 to round 2 is illustrated in Fig. 5 and Table 6 in the Supplementary File, and confirms that reliability of the raters was consistent across both rounds, independent of the number of raters.

The calibration session effectively addressed variations, however, it is unclear if a single 2-hour session would ensure long-term

improvement in how consistently raters assess. Four participant videos were reviewed, which may not fully represent the diversity of cases, and rating a video may not be the same as rating face-to-face in real-time.

6. Future works

Clinicians' agreement improved after a training session, and relatedness between hands decreased, indicating an ability to focus on each side separately. Future work could target different training modalities and timeframes to examine the possibility of maintaining agreement between experienced clinicians over time. Scaling up of clinician PD training across disciplines would require careful planning and collaboration. Evaluator training, experience, and commitment to improving assessment quality are key factors influencing inter-rater evaluation [47]. Any future implementation of a training intervention for PD assessment should asses the long-term retention of the training effects and the scalability across diverse healthcare settings (e.g. from specialised centres to primary care) and across clinician experience levels. As research shows that knowledge decay can occur without reinforcement strategies [48], follow-up refresher courses may be necessary for sustained effectiveness. There is also scope to further explore the role of technology to improve both PD assessment reliability and overall care quality, including its acceptability and feasibility in busy clinic settings.

7. Conclusion

This study demonstrates significant variability in ratings between experienced clinicians when assessing video-recorded hand movements. This can be mitigated to some extent with a training and calibration session. Although improvements were observed, even after training, variability in ratings persisted and suggests that assessment of motor features remains complex, emphasising the importance of improving consistency of clinical assessments in PD, including the role of technology-enhanced assessments.

8. Ethical compliance statement

Ethical approval was obtained from the Clinical Research Ethics Committee of the University College, Cork, ECM 4 (a) 16/10/19.

Informed written consent was obtained from participants for their involvement in the research and for access to their personal data. Participants were provided with information leaflets detailing the study objectives, had the opportunity to pose questions, and subsequently signed a consent form, with both participant and researcher retaining a copy.

9. Funding statement

This work was supported by the Interreg Northern Periphery and Artic Programme funded project SENDOC (Smart sENsor Devices fOr rehabilitation and Connected health).

CRediT authorship contribution statement

Lorna Kenny: Writing – review & editing, Writing – original draft, Visualization, Validation, Methodology, Investigation, Formal analysis, Data curation. **Zahra Azizi:** Writing – review & editing, Writing – original draft, Visualization, Validation, Formal analysis, Data curation. **Kevin Moore:** Writing – review & editing, Methodology, Data curation. **Megan Alcock:** Writing – review & editing, Methodology, Investigation. **Sarah Heywood:** Writing – review & editing, Methodology, Investigation. **Agnes Jonsson:** Writing – review & editing, Methodology, Investigation. **Keith McGrath:** Writing – review & editing, Methodology, Investigation. **Mary J. Foley:** Writing – review & editing, Methodology, Investigation. **Brian Sweeney:** Writing – review & editing, Methodology, Investigation. **Sean O'Sullivan:** Writing – review & editing, Methodology, Investigation. **John Barton:** Writing – review & editing, Software, Methodology, Funding acquisition. **Salvatore Tedesco:** Writing – review & editing, Software, Methodology. **Marco Sica:** Writing – review & editing, Methodology, Investigation. **Colum Crowe:** Writing – review & editing, Methodology, Investigation. **Suzanne Timmons:** Writing – review & editing, Writing – original draft, Supervision, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

The authors would like to thank Parkinson's Ireland and its members for their support and guidance in this study. Additionally, our appreciation is extended to all participants who generously volunteered their time and effort. Thank you to our funders Interreg Northern Periphery and Artic Programme funded project SENDOC (Smart sENsor Devices fOr rehabilitation and Connected health).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.prdoa.2024.100278.

References

[1] E.R. Dorsey, B.R. Bloem, The Parkinson Pandemic—A Call to Action, JAMA Neurol. 75 (1) (2018 Jan 1) 9, https://doi.org/10.1001/jamaneurol.2017.3299.
[2] Ou Z, Pan J, Tang S, Duan D, Yu D, Nong H, et al. Global Trends in the Incidence, Prevalence, and Years Lived With Disability of Parkinson's Disease in 204 Countries/Territories From 1990 to 2019. Front Public Health [Internet]. 2021 [cited 2023 Dec 21];9. Available from: https://www.frontiersin.org/articles/10.3389/fpubh.2021.776847.
[3] R.B. Postuma, D. Berg, M. Stern, W. Poewe, C.W. Olanow, W. Oertel, et al., MDS clinical diagnostic criteria for Parkinson's disease: MDS-PD Clinical Diagnostic Criteria, Mov. Disord. 30 (12) (2015 Oct) 1591–1601, https://doi.org/10.1002/mds.26424.
[4] C.G. Goetz, B.C. Tilley, S.R. Shaftman, G.T. Stebbins, S. Fahn, P. Martinez-Martin, et al., Movement Disorder Society-sponsored revision of the Unified Parkinson's Disease Rating Scale (MDS-UPDRS): Scale presentation and clinimetric testing results, Mov. Disord. 23 (15) (2008 Nov 15) 2129–2170, https://doi.org/10.1002/mds.22340.
[5] Y. Guo, G.T. Stebbins, T.A. Mestre, C.G. Goetz, S. Luo, Movement Disorder Society Unified Parkinson's Disease Rating Scale Motor Examination Retains Its 2-Domain Profile in Both On and Off States, Mov Disord Clin Pract. 9 (8) (2022 Nov) 1149–1151, https://doi.org/10.1002/mdc3.13566.
[6] B. Post, M.P. Merkus, R.M.A. De Bie, R.J. De Haan, J.D. Speelman, Unified Parkinson's disease rating scale motor examination: Are ratings of nurses, residents in neurology, and movement disorders specialists interchangeable? Mov. Disord. 20 (12) (2005 Dec) 1577–1584, https://doi.org/10.1002/mds.20640.
[7] Lange RT. Inter-rater Reliability. In: Kreutzer JS, DeLuca J, Caplan B, editors. Encyclopedia of Clinical Neuropsychology [Internet]. New York, NY: Springer; 2011 [cited 2023 Dec 21]. p. 1348–1348. Available from: https://doi.org/10.1007/978-3-319-56782-2_1203-2.
[8] A. Regnault, B. Boroojerdi, J. Meunier, M. Bani, T. Morel, S. Cano, Does the MDS-UPDRS provide the precision to assess progression in early Parkinson's disease? Learnings from the Parkinson's progression marker initiative cohort, J. Neurol. 266 (8) (2019 Aug) 1927–1936, https://doi.org/10.1007/s00415-019-09348-3.
[9] L.J.W. Evers, J.H. Krijthe, M.J. Meinders, B.R. Bloem, T.M. Heskes, Measuring Parkinson's disease over time: The real-world within-subject reliability of the MDS-UPDRS, Mov Disord off J Mov Disord Soc. 34 (10) (2019 Oct) 1480–1487, https://doi.org/10.1002/mds.27790.
[10] C.H. Adler, T.G. Beach, N. Zhang, H.A. Shill, E. Driver-Dunckley, S.H. Mehta, et al., Clinical Diagnostic Accuracy of Early/Advanced Parkinson Disease, Neurol Clin Pract. 11 (4) (2021 Aug) e414–e421, https://doi.org/10.1212/CPJ.0000000000001016.
[11] Falup-Pecurariu C, Ferreira J, Martinez-Martin P, Chaudhuri KR, editors. Movement Disorders Curricula [Internet]. Vienna: Springer; 2017 [cited 2023 Dec 21]. Available from: http://link.springer.com/10.1007/978-3-7091-1628-9.
[12] Deb R, Bhat G, An S, Shill H, Ogras UY. Trends in Technology Usage for Parkinson's Disease Assessment: A Systematic Review [Internet]. medRxiv; 2021 [cited 2024 Oct 4]. p. 2021.02.01.21250939. Available from: https://www.medrxiv.org/content/10.1101/2021.02.01.21250939v1.

[13] Intelligent Sensory Pen for Aiding in the Diagnosis of Parkinson's Disease from Dynamic Handwriting Analysis [Internet]. [cited 2024 Oct 4]. Available from: https://www.mdpi.com/1424-8220/20/20/5840.
[14] T. Fay-Karmon, N. Galor, B. Heimler, A. Zilka, R.P. Bartsch, M. Plotnik, et al., Home-based monitoring of persons with advanced Parkinson's disease using smartwatch-smartphone technology, Sci. Rep. 14 (1) (2024 Jan 2) 9, https://doi.org/10.1038/s41598-023-48209-y.
[15] T. Exley, S. Moudy, R.M. Patterson, J. Kim, M.V. Albert, Predicting UPDRS Motor Symptoms in Individuals With Parkinson's Disease From Force Plates Using Machine Learning, IEEE J. Biomed. Health Inform. 26 (7) (2022 Jul) 3486–3494, https://doi.org/10.1109/JBHI.2022.3157518.
[16] C. Herbers, R. Zhang, A. Erdman, M.D. Johnson, Distinguishing features of Parkinson's disease fallers based on wireless insole plantar pressure monitoring, NPJ Park Dis. 19 (10) (2024 Mar) 67, https://doi.org/10.1038/s41531-024-00678-2.
[17] Sibley KG, Girges C, Hoque E, Foltynie T. Video-Based Analyses of Parkinson's Disease Severity: A Brief Review. J Park Dis. 11(Suppl 1):S83–93. https://doi.org/10.3233/JPD-202402.
[18] D. Rudå, G. Einarsson, A.S.S. Andersen, J.B. Matthiassen, C.U. Correll, K. Winge, et al., Exploring Movement Impairments in Patients With Parkinson's Disease Using the Microsoft Kinect Sensor: A Feasibility Study, Front. Neurol. 6 (11) (2021 Jan) 610614, https://doi.org/10.3389/fneur.2020.610614.
[19] S.K. Khare, V. Bajaj, U.R. Acharya, Detection of Parkinson's disease using automated tunable Q wavelet transform technique with EEG signals, Biocybern Biomed Eng. 41 (2) (2021 Apr 1) 679–689, https://doi.org/10.1016/j.bbe.2021.04.008.
[20] C. Moreau, T. Rouaud, D. Grabli, I. Benatru, P. Remy, A.R. Marques, et al., Overview on wearable sensors for the management of Parkinson's disease, Npj Park Dis. 9 (1) (2023 Nov 2) 1–16, https://doi.org/10.1038/s41531-023-00585-y.
[21] L. Kenny, K. Moore, C. O' Riordan, S. Fox, J. Barton, S. Tedesco, et al., The Views and Needs of People With Parkinson Disease Regarding Wearable Devices for Disease Monitoring: Mixed Methods Exploration, JMIR Form Res. (2022), https://doi.org/10.2196/27418. Jan 6;6(1):e27418.
[22] A.J. Espay, P. Bonato, F.B. Nahab, W. Maetzler, J.M. Dean, J. Klucken, et al., Technology in Parkinson's disease: challenges and opportunities, Mov Disord off J Mov Disord Soc. 31 (9) (2016 Sep) 1272–1282, https://doi.org/10.1002/mds.26642.
[23] C.G. Goetz, G.T. Stebbins, Assuring interrater reliability for the UPDRS motor section: Utility of the UPDRS teaching tape, Mov. Disord. 19 (12) (2004) 1453–1456, https://doi.org/10.1002/mds.20220.
[24] M. Imran, T.F. Halawa, M. Baig, A.M. Almanjoumi, M.M. Badri, W.A. Alghamdi, Team-based learning versus interactive lecture in achieving learning outcomes and improving clinical reasoning skills: a randomized crossover study, BMC Med. Educ. 22 (1) (2022 May 7) 348, https://doi.org/10.1186/s12909-022-03411-w.
[25] M. Richards, K. Marder, L. Cote, R. Mayeux, Interrater reliability of the Unified Parkinson's Disease Rating Scale motor examination, Mov Disord off J Mov Disord Soc. 9 (1) (1994 Jan) 89–91, https://doi.org/10.1002/mds.870090114.
[26] F.T. De Deus, D. Santos García, A.M. Macías, Variabilidad en la exploración motora de la enfermedad de Parkinson entre el neurólogo experto en trastornos del movimiento y la enfermera especializada, Neurología. 34 (8) (2019 Oct) 520–526, https://doi.org/10.1016/j.nrl.2017.03.005.
[27] Kremer NI, Smid A, Lange SF, Marçal IM, Tamasi K, Dijk JMC van, et al. Supine MDS-UPDRS-III Assessment: An Explorative Study. J Clin Med. 12(9):3108. https://doi.org/10.3390/jcm12093108.
[28] N. Wendel, C.E. Macpherson, K. Webber, K. Hendron, T. DeAngelis, C. Colon-Semenza, et al., Accuracy of Activity Trackers in Parkinson Disease: Should We Prescribe Them? Phys. Ther. 98 (8) (2018 Aug 1) 705–714, https://doi.org/10.1093/ptj/pzy054.
[29] T.K. Koo, M.Y. Li, A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research, J. Chiropr. Med. 15 (2) (2016 Jun) 155–163, https://doi.org/10.1016/j.jcm.2016.02.012.
[30] G.T. Stebbins, C.G. Goetz, Factor structure of the Unified Parkinson's Disease Rating Scale: Motor Examination section, Mov Disord off J Mov Disord Soc. 13 (4) (1998 Jul) 633–636, https://doi.org/10.1002/mds.870130404.
[31] S. Williams, D. Wong, J.E. Alty, S.D. Relton, Parkinsonian Hand or Clinician's Eye? Finger Tap Bradykinesia Interrater Reliability for 21 Movement Disorder Experts, J Park Dis. 13 (4) (2023 Jun 13) 525–536, https://doi.org/10.3233/JPD-223256.
[32] G. Morinan, Y. Dushin, G. Sarapata, S. Rupprechter, Y. Peng, C. Girges, et al., Computer vision quantification of whole-body Parkinsonian bradykinesia using a large multi-site population, Npj Park Dis. 9 (1) (2023 Jan 27) 1–12, https://doi.org/10.1038/s41531-023-00454-8.
[33] M. Amer, G. Hubert, S.J. Sullivan, P. Herbison, E.A. Franz, G.D. Hammond-Tooke, Reliability and diagnostic characteristics of clinical tests of upper limb motor function, J Clin Neurosci off J Neurosurg Soc Australas. 19 (9) (2012 Sep) 1246–1251, https://doi.org/10.1016/j.jocn.2011.12.007.
[34] D.A. Heldman, J.P. Giuffrida, R. Chen, M. Payne, F. Mazzella, A.P. Duker, et al., The modified bradykinesia rating scale for Parkinson's disease: reliability and comparison with kinematic measures, Mov Disord off J Mov Disord Soc. 26 (10) (2011 Aug 15) 1859–1863, https://doi.org/10.1002/mds.23740.
[35] E. Heinrichs-Graham, P.M. Santamaria, H.E. Gendelman, T.W. Wilson, The cortical signature of symptom laterality in Parkinson's disease, NeuroImage Clin. 12 (14) (2017 Feb) 433–440, https://doi.org/10.1016/j.nicl.2017.02.010.
[36] C.R. Baumann, U. Held, P.O. Valko, M. Wienecke, D. Waldvogel, Body side and predominant motor features at the onset of Parkinson's disease are linked to motor and nonmotor progression, Mov Disord off J Mov Disord Soc. 29 (2) (2014 Feb) 207–213, https://doi.org/10.1002/mds.25650.
[37] Lahr J, Pereira MP, Pelicioni PHS, De Morais LC, Gobbi LTB. Parkinson's Disease Patients with Dominant Hemibody Affected By The Disease Rely More On Vision To Maintain Upright Postural Control. Percept Mot Skills. 2015 Dec;121(3):923–34. https://doi.org/ 10.2466/15.PMS.121c26x0.
[38] J. Ren, C. Pan, Y. Li, L. Li, P. Hua, L. Xu, et al., Consistency and Stability of Motor Subtype Classifications in Patients with de novo Parkinson's Disease, Front. Neurosci. 1 (15) (2021 Mar) 637896, https://doi.org/10.3389/fnins.2021.637896.
[39] D.A. Delgado, B.S. Lambert, N. Boutris, P.C. McCulloch, A.B. Robbins, M. R. Moreno, et al., Validation of Digital Visual Analog Scale Pain Scoring with a Traditional Paper-based Visual Analog Scale in Adults, J Am Acad Orthop Surg Glob Res Rev. 2 (3) (2018 Mar 23) e088.
[40] Perez-Lloret S, Ciampi de Andrade D, Lyons KE, Rodríguez-Blázquez C, Chaudhuri KR, Deuschl G, et al. Rating Scales for Pain in Parkinson's Disease: Critique and Recommendations. Mov Disord Clin Pract. 2016 Jun 24;3(6):527–37. https://doi.org/10.1002/mdc3.12384.
[41] P. Martínez-Martín, J.M. Rojo-Abuin, M. Rodríguez-Violante, M. Serrano-Dueñas, N. Garretto, J.C. Martínez-Castrillo, et al., Analysis of four scales for global severity evaluation in Parkinson's disease, NPJ Park Dis. 5 (2) (2016 May) 16007, https://doi.org/10.1038/npjparkd.2016.7.
[42] EuroQol Group, EuroQol–a new facility for the measurement of health-related quality of life, Health Policy Amst Neth. 16 (3) (1990 Dec) 199–208, https://doi.org/10.1016/0168-8510(90)90421-9.
[43] C.G. Goetz, W. Poewe, O. Rascol, C. Sampaio, G.T. Stebbins, C. Counsell, et al., *Movement* Disorder Society Task Force report on the Hoehn and Yahr staging scale: Status and recommendations The *Movement* Disorder Society Task Force on rating scales for Parkinson's disease, Mov. Disord. 19 (9) (2004 Sep) 1020–1028, https://doi.org/10.1002/mds.20213.
[44] S.D. Aradi, R.A. Hauser, Medical Management and Prevention of Motor Complications in Parkinson's Disease, Neurotherapeutics 17 (4) (2020 Oct 1) 1339–1365, https://doi.org/10.1007/s13311-020-00889-4.
[45] C.A. Artusi, L. Lopiano, F. Morgante, Deep Brain Stimulation Selection Criteria for Parkinson's Disease: Time to Go beyond CAPSIT-PD, J. Clin. Med. 9 (12) (2020 Dec 4) 3931, https://doi.org/10.3390/jcm9123931.
[46] P.A. Vaswani, T.F. Tropea, N. Dahodwala, Overcoming Barriers to Parkinson Disease Trial Participation: Increasing Diversity and Novel Designs for Recruitment and Retention, Neurotherapeutics 17 (4) (2020 Oct) 1724–1735, https://doi.org/10.1007/s13311-020-00960-0.
[47] B.RC. Shweta, H.K. Chaturvedi, Evaluation of inter-rater agreement and inter-rater reliability for observational data: An overview of concepts and methods, J. Indian Acad. Appl. Psychol. 41 (3) (2015) 20–27.
[48] Parkinson's UK [Internet]. [cited 2024 Oct 1]. Explore the Parkinson's learning pathway for health and care staff. Available from: https://www.parkinsons.org.uk/professionals/parkinsons-learning-pathway-health-care-staff.