Original Research Article

SAGE Open Medicine

Reliability of functional outcome measures in adults with neurofibromatosis 2

SAGE Open Medicine
Volume 10: 1–9

DOI: 10.1177/20503121221118996
journals.sagepub.com/home/smo
SAGE

Rebecca Louise Mullin[1,2], Rebecca Smith[1,2], Susan Wood[1], Angela Swampillai[1] and Shazia Afridi[1]

Abstract

Objective: To determine inter- and intra-rater reliability of functional performance outcome measures in people with neurofibromatosis 2. To ascertain how closely objective and subjective measures align.

Methods: Twenty-nine people with neurofibromatosis 2 were recorded performing the modified clinical test of sensory integration and balance, four square step test and modified nine-hole peg tests. Three raters scored each measure to determine inter-rater reliability. One rater scored the measures a second time to determine intra-rater reliability. Participants also completed a disease-specific quality of life questionnaire and dynamic visual acuity testing.

Results: Inter-rater and intra-rater reliability scores (intra-class correlation coefficient) were excellent for all tests (intra-class correlation coefficient $r \geqslant 0.9$). The four square step test correlated with perceived walking challenges and modified clinical test of sensory integration and balance correlated with perceived balance challenges in a neurofibromatosis 2 quality of life patient report outcome measure.

Conclusion: The modified clinical test of sensory integration and balance, four square step test and modified nine-hole peg tests are potentially useful measures for monitoring neurofibromatosis 2.

Keywords

Neurofibromatosis 2, NF2, reliability, outcome measures, standard error of measurement, minimal detectable change

Date received: 17 March 2022; accepted: 22 July 2022

Introduction

Neurofibromatosis type 2 (NF2) is a rare, inherited disease which can lead to functional challenges.[1] NF2 is characterised by vestibular schwannomas (VS) and schwannomas may also form on other cranial, spinal and peripheral nerves. Meningiomas and ependymomas as well as neuropathies may also develop.[2] Balance difficulties, hearing loss, visual disturbance and impairments in dexterity are functional challenges affecting people with NF2.

The advent of bevacizumab treatment has led to decreased VS growth rates, improved hearing and quality of life (QOL) in NF2.[3] Anecdotally, it also appears that functional challenges may improve with bevacizumab treatment. However, to date there is no way of confidently quantifying changes in functional task performance in this disease group as no outcome measures have undergone sufficient psychometric testing, and if used, there is an inherent element of doubt in interpreting the results. Often the subjective experience of balance, walking and hand dexterity challenges do not align with the clinical examination or investigations, presenting a challenge to both the clinician and patient.

An essential requirement of all outcome measurements is that they are reliable.[4] Reliability is defined as 'the degree to which measurement is free from measurement error'[5]. The observed score of an outcome measure is a composite of the true score and random error which may occur at any point in the measurement process as a result of fatigue, inattention or inaccuracy.[4] Inter-rater reliability requires the same group of subjects to be measured at the same time by different observers while intra-rater reliability considers the same subjects and the same observer with measurements taken at different time points.[6] Absolute reliability is expressed as the standard error of measurement (SEM), and this can be calculated from

[1]National Centre for Neurofibromatosis, Guy's and St Thomas' NHS Foundation Trust, London, UK
[2]Department of Physiotherapy, Guy's and St Thomas' NHS Foundation Trust, London, UK

Corresponding author:
Rebecca Louise Mullin, Department of Physiotherapy, Guy's and St Thomas' NHS Foundation Trust, Third Floor Lambeth Wing, St Thomas' Hospital, London SE1 7EH, UK.
Email: Rebecca.mullin@gstt.nhs.uk

intra-rater reliability. Minimal detectable change (MDC) describes the minimal amount of change in the instrument score to determine that the score change is not attributable to measurement error and this may be calculated from SEM.[7] It is important to evaluate properties such as reliability of an outcome measure within the target population, as variability within that condition strongly influences outcome measurement results.[4]

Bilateral VS, a characteristic feature of NF2, impairs vestibular function and more specifically the vestibular ocular reflex (VOR).[8] The primary function of the VOR is to stabilise gaze when the head is moving. Thus, patients with bilateral VOR impairment experience balance difficulties and oscillopsia during head movements, making functional tasks challenging.[9] Dynamic visual acuity (DVA) testing evaluates the VOR within a functional context and has been examined for reliability, sensitivity and specificity in patients with unilateral and bilateral vestibular loss.[10] Although evidence shows DVA is significantly impaired in patients with bilateral vestibular loss.[10] To the best of our knowledge, it has not been specifically evaluated in people with NF2. It is thought DVA testing may be a useful adjunct in monitoring the impact of VS on patients' balance and as a way or measuring rehabilitation outcomes in the NF2 population. Indeed, DVA has been evaluated[11] and used as an outcome measure following vestibular rehabilitation in patients with unilateral vestibular hypofunction.[9]

The primary aim of this study is to evaluate inter- and intra-rater reliability of three commonly used outcome measures in adults who have NF2. From these data, we will calculate the SEM and MDC. We also aim to ascertain how closely performance of the chosen functional outcome measures correlates with subjective experience of balance in a disease-specific, QOL questionnaire and self-report falls history. A secondary aim of this study is to collect DVA data for people with NF2.

Methods

Guy's and St Thomas' NHS Foundation Trust is a national centre for the diagnosis, management and support of people with NF2. One-hundred fifty-five adults (aged 16 years and over) with NF2 who attended their clinic appointments during the 7-month recruitment period were approached by letter with participant information sheet, inviting them to take part in this observational study. We aimed to recruit 50 participants as recommended for a reliability study,[4] as this type of study does not require a power calculation,[4] while recognising that NF2 is a rare disease. At the time of their appointment, the treating clinician (doctor or nurse) confirmed that they met the inclusion/exclusion criteria and ascertained whether they wished to take part. If they volunteered, they were introduced to a researcher (R.S./R.L.M.).

To be included in this study, the participants needed to meet the following requirements: have a clinical diagnosis of NF2, be aged 16 years or over, attend the National NF service at Guy's and St Thomas' NHS Foundation Trust, be able to provide informed consent, not have significant mobility or balance impairments that are unrelated to their NF2 and be able to walk more than 10 m without physical assistance (may use walking aids). Exclusion criteria included having an unstable vascular or orthopaedic pathology of the cervical spine or having had a stroke within the past 3 months.

After being given the opportunity to ask questions about the study, written consent was collected by a researcher (R.S./R.L.M.) and the participant was given a unique alphanumeric research identification code. All study participants (100%) provided demographic information and completed a neurofibromatosis 2 quality of life patient report outcome measure (NFTI-QOL). They also provided a falls history including whether or not they had fallen over or had any 'near misses' in the previous year. Each participant was asked to complete three repetitions of each of the chosen outcome measures while being video recorded by the researcher (R.S./R.L.M.) and were given time to rest between trials as required. They then completed the DVA test (Figure 1).

Outcome measure selection. Outcome measures were selected following a review of the evidence base and pre-study consultations with stake holders.

The evidence base was reviewed to identify functional performance outcome measures which have undergone metric evaluation in comparable cohorts including those with vestibular pathology, community dwelling adults with balance deficits and people with neuropathy.

Pre-study consultations involved stakeholders (the treating clinical team and patients), selected to ensure measures were clinically useful and pertinent to patients.[12] Patients ($n = 10$) were asked to first identify their most problematic symptoms and second to rank their symptoms in order (most problematic to least problematic). Analysis of these data demonstrated patients ranked balance, hearing, blurry vision and upper limb function as problematic symptoms. As standardised measures of audition are already utilised within NF2, it was deemed appropriate to focus on selecting outcome measures which would evaluate balance, blurry vision and upper limb function.

Further discussions with stakeholders regarding criteria for long-term outcome measure use (i.e. the measure being easy and quick to carry out by different members of the multidisciplinary team (MDT) and the findings being quick and easy to interpret) led to selection of four outcome measures related to the impairments identified by patients. Two outcome measures were selected pertaining to balance (examining static and dynamic balance, respectively) – modified clinical test of sensory interaction in balance (mCTSIB) and four square step test (FSST). One measure was selected to explore upper limb function (modified nine-hole peg test (m9HPT)) and a further measure to quantify blurry vision – DVA. Further details of each measure are shown in the following sections.

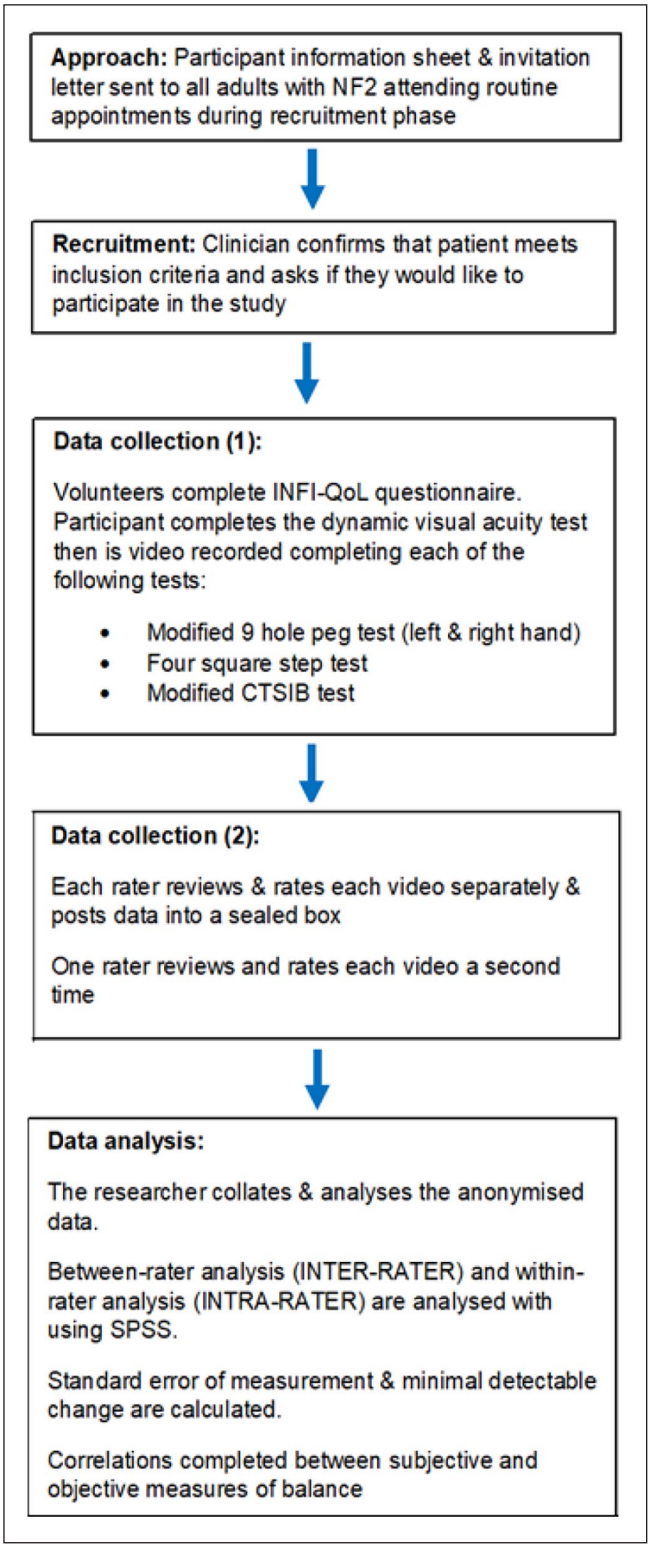
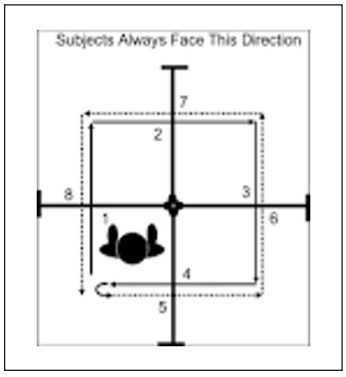
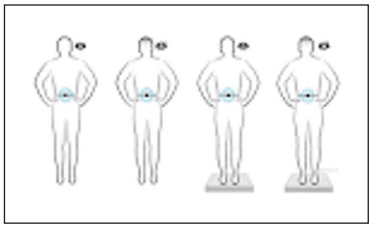


Figure 1. Study flow diagram.

FSST. The FSST assesses dynamic balance.[13] In the FSST, the participant steps forwards, to the side, backwards, then to the side in a square and then reverses. Time is recorded to the nearest millisecond. A longer time taken to complete the test indicates worse balance. The FSST has been validated in a variety of neurological populations.[14]



Modified clinical test of sensory integration and balance. The modified clinical test of sensory integration and balance (mCTSIB) assesses static balance in four different conditions.[15] Each stage is completed if the position can be held for 30 s. A total sum score is recorded out of 120 (s) with a higher score indicating better balance:



- Condition 1: feet together, eyes OPEN, STABLE surface.
- Condition 2: feet together, eyes CLOSED, STABLE surface.
- Condition 3: feet together, eyes OPEN, UNSTABLE surface.
- Condition 4: feet together, eyes CLOSED, UNSTABLE surface.

M9HPT. The m9HPT assesses upper limb function through dexterity. In this test, the participant takes pegs from a bowl and places them into the holes of a peg board.[16] A higher score indicates worse dexterity. Measurements are recorded to milliseconds.

DVA test. The DVA test is administered by first testing the participant's static visual acuity. The participant is seated a specified distance from the computer screen and is instructed to wear their usual prescription lenses if required. With the head still, the participant is instructed to read letters from the computer screen until they can no longer complete the task. The sequence is performed again, this time with the head passively moved in the yaw plane at 2 Hz, or two cycles per second at a magnitude of 20° to 30° from the midline. Letters are randomly presented during the task to negate a learned effect. A DVA score is generated by comparing the two conditions. A score of >0.2 LogMAR is considered abnormal.[17]

Subjective symptom reporting (NFTI-QOL questionnaire). All participants (29/29) completed the NFTI-QOL questionnaire in order to ascertain how closely subjective experiences of symptoms correlated with objective performance on outcome measures. The NFTI-QOL questionnaire[18] is a validated, reliable disease-specific QOL self-report outcome measure. The NFTI-QOL was developed within Guy's and St Thomas' NHS Foundation Trust, and permission was granted for use within this study. Responders categorise problems as no issues, mild, moderate or severe in eight domains. It includes two functional questions where responders rate their challenges of walking and balance.

Rating process. Video recordings of participants performing the outcome measurement tests were immediately transferred to a secure electronic location. Three raters watched and rated the videos separately to assess inter-rater reliability, and they posted their scores for each test into a sealed box. One of the raters, rated the video a second time, to assess intra-rater reliability. The rater team comprised of three experienced members of the NF2 MDT including doctors and a clinical nurse specialist. The researcher collated these data onto a spreadsheet. Data were transferred to SPSS for statistical analysis.

Bias. Several steps were taken to counter bias in this study. The researcher was not involved in the recruitment processes or as a video rater to reduce the risk of selection bias. The intra-rater reliability tester was instructed to watch the videos a second time, only after they had watched all 29 sets of videos through once to ameliorate recall bias. Outcome measurements were completed with two researchers (R.L.M. and R.S.) with standardised instructions to reduce the risk of performance bias. Videos were taken of the outcome measurement sessions and used for analysis to ensure all raters saw the same test, from the same angle after receiving training on how to score the tests to reduce the risk of detection bias.

Statistical analysis. A statistician supported the research team throughout the research process. Data from all measures were analysed using the IBM Statistical Package for Social Sciences (SPSS) version 23. A two-way mixed effects model was used to calculate intra-class correlation coefficient (ICC 3,1) and evaluate relative intra-rater reliability of the mCTSIB, FSST and m9HPT. A two-way random effects model was used (ICC 2,1) to evaluate inter-rater reliability of the mCTSIB, FSST and m9HPT (Table 2). The statistical analyses align with other studies investigating inter- and intra-rater reliability of the selected functional outcome measures.[19–21]

The ICC is a number between zero and one – one represents the perfect reliability with no measurement error, and zero represents no reliability. Values less than 0.5 are indicative of poor reliability, values between 0.5 and 0.75 indicate the moderate reliability, values between 0.75 and 0.9 indicate the good reliability, and values greater than 0.90 indicate the excellent reliability.[22]

The SEM, absolute inter- and intra-rater reliability, was calculated for each measure in adults with NF2 with the following equation

$$\mathrm{SEM} = \mathrm{SD}\sqrt{1-r}$$

MDC was calculated for each measure from the SEM using the following equation

$$\mathrm{MDC} = \mathrm{SEM} \times 1.96 \times \sqrt{2}$$

In the above equations, MDC is the minimal detectable change, SEM is the standard error of measurement, and r is the reliability (ICC).

Results

A total of 45 adults with NF2 were invited to take part in this study. Four did not receive the information sheet before their appointment (more than 24 h), five left the department before being seen by the researcher, four did not consent to take part and three did not meet the inclusion criteria. Thirty were recruited into the study. One participant died during the recruitment period and in line with our ethical agreement, their data were removed before analysis. Subsequently, 29 sets of participant data underwent statistical analysis.

Table 1 details that there were 18 females and 11 males in this study. Median age was 45.5 years (range = 18–73). Thirty four percent were on the bevacizumab treatment. Participant presentations were varied as seen in Table 4. The range of scores for the NFTI-QOL questionnaire was 2–19 and the mean value was 9, where 0 indicates no difficulty and 24 indicates severe difficulty in all eight domains.

Table 2 details the ICC for intra-rater and inter-rater reliability for each of the three outcome measurements with 95% confidence intervals (95% CIs). Reliability (ICC) was excellent with low measurement error and tight 95% CIs for all outcome measures, and intra-rater reliability findings were comparable with inter-rater reliability findings.

Table 3 details the mean score and range for each of the above tests alongside clinically important MDC. There was

Table 1. Demographic details of participants.

Participant	Age (years)	Gender	Vestibular schwannoma	Spinal tumours	Peripheral neuropathy	Visual disturbance	On bevacizumab treatment
AO1012	29	M	Bilateral	Yes	No	No	Yes
A11125	27	F	Bilateral	Yes	Yes	No	Yes
A31159	37	F	Bilateral	Yes	Yes	Yes	Yes
A40917	37	M	Bilateral	Yes	Yes	Yes	Yes
A51045	62	F	Bilateral	No	No	No	No
A61057	28	F	Bilateral	No	No	No	No
A71114	34	F	Bilateral	Yes	Yes	Yes	Yes
A81438	32	F	Bilateral	Yes	No	Yes	No
A91046	18	F	Bilateral	Yes	No	Yes	No
A101049	44	M	Bilateral	Yes	No	No	No
A111123	37	F	Bilateral	Yes	No	No	Yes
A121200	27	F	Unilateral	Yes	No	No	Yes
A131313	25	M	Bilateral	Yes	Yes	No	No
A141015	23	F	Bilateral	Yes	Yes	Yes	Yes
A151154	57	F	Unilateral	Yes	No	No	No
A161000	43	M	Bilateral	Yes	Yes	No	No
A171058	73	M	Bilateral	Yes	Yes	No	No
A181133	64	F	Unilateral	No	Yes	Yes	No
A191027	19	M	Bilateral	Yes	Yes	Yes	Yes
A201109	26	F	Bilateral	Yes	No	Yes	Yes
A211044	36	F	Bilateral	Yes	No	No	No
A221142	64	F	Bilateral	Yes	No	No	No
A231118	45	F	Bilateral	No	No	No	No
A241157	47	F	Unilateral	Yes	No	No	No
A251345	32	M	Bilateral	Yes	Yes	Yes	No
A261134	69	M	Bilateral	Yes	No	Yes	No
A271145	54	F	Unilateral	Yes	No	No	No
A281239	26	M	Bilateral	Yes	No	No	No
A281239	24	M	Bilateral	No	No	No	No

M: male; F: female.

Table 2. Rater reliability, intra-class correlation coefficient (ICC) scores with 95% confidence intervals (95% CI) for outcome measures.

Functional test	Intra-rater reliability		Inter-rater reliability	
	ICC	95% CI	ICC	95% CI
mCTSIB	0.999	0.997–0.999	0.999	0.997–0.999
FSST	0.998	0.995–0.999	0.998	0.995–0.999
m9HPT	0.999	0.999–1.00	0.999	0.999–1.00

mCTSIB: modified clinical test of sensory integration and balance; FSST: four square step test; m9HPT: modified nine-hole peg test.

a normal distribution for all measurement results. The wide range of times was not simply due to outliers but probably reflected the clinical heterogeneity of NF2 and the participant's level of perceived functional difficulty.

Table 4 details the correlations between subjective self-report experience and objective functional performance in the two balance and mobility outcome measures. The first three columns represent correlations with the NFTI-QOL questionnaire which measured patient reported, disease-specific QOL. Pearson's correlations were computed between each functional test and the total NFTI-QOL questionnaire scores, subsections for question 1 for balance and question 5 for walking.

Correlations were the strongest between the modified CTSIB and questions related to balance in the NFTI-QOL, and the FSST with NFTI-QOL total scores. A weaker but statistically significant correlation was also found between the FSST and question 5 for walking. Table 3 also details correlations with participants self-report falls history. Participants were subcategorised based on their falls history

Table 3. Mean scores for each outcome measure with standard deviation, range, standard error of measurement and minimal detectable change scores, calculated from inter-rater reliability.

Functional test units	Mean score	1 Standard deviation	Range (minimum to maximum)	Standard error of measurement (SEM)	Minimal detectable change (MDC)
mCTSIB total score (/120) (s)	87.35	21.42	35.8–120	3.48	9.66
mCTSIB condition 1 (/30) (s)	30	0	30	n/a	n/a
mCTSIB condition 2 (/30) (s)	27.33	7.86	0–30	1.15	3.19
mCTSIB condition 3 (/30) (s)	29.43	2.95	15–30	0.37	0.1
mCTSIB condition 4 (/30) (s)	8.41	11.06	0–30	0.14	0.38
FSST (s)	10.54	2.64	6.13–17.72	0.21	0.58
m9HPT (left) (s)	19.50	8.21	10.19–51.04	0.38	1.05
m9HPT (right) (s)	18.55	5.77	10.29–40.09	0.33	0.84

mCTSIB: modified clinical test of sensory integration and balance; FSST: four square step test; m9HPT: modified nine-hole peg test; n/a: not available.

Table 4. Correlation between subjective self-report and objective balance outcome measures.

Functional test	Correlation with NFTI-QOL total (*r*)	Correlation with NFTI-QOL balance (*r*)	Correlation with NFTI-QOL walking (*r*)	Correlation with self-report falls history
FSST	0.527**	0.323 (ns)	0.407*	0.198 (ns)
mCTSIB	0.36*	0.48**	0.32*	−0.417*

NFTI-QOL: neurofibromatosis 2 quality of life patient report outcome measure; FSST: four square step test; mCTSIB: modified clinical test of sensory integration and balance; ns: not significant.
Pearson's (*r*) between functional test and subsections of the NFTI-QOL with significance level. Correlation with self-report falls history using Scheffe's post hoc ANOVA analysis.
*$p < 0.05$; **$p < 0.01$.

in the last year into one of three categories: 'no falls or near misses' (4/29), 'near misses but no falls' (12/29) or 'falls with or without near misses' (13/29). 'No falls but near misses' and 'no falls or near misses' could compress to 'non-fallers' (16/29) for comparison with the 'fallers' group (13/29) as required. Analyses of variance (ANOVAs) compared the mean values between these three categories and a post hoc Scheffe's test analysed significance. There was a clear trend of increasing mean scores in the FSST as falls history worsened, although this did not reach statistical significance and it did not discriminate between 'fallers' and 'non-fallers'. For the modified CTSIB, mean values for the two groups who represented non-fallers – 'no falls or near misses' and 'near misses but no falls' – were comparable and not significantly different. However, the 'fallers' group has lower total mCTSIB scores than the 'non-fallers', and this reached statistical significance, discriminating 'fallers' from 'non-fallers' with a total sum score of less than 88 s.

Twenty-four out of twenty-nine participants completed the DVA test due to technical problems. Mean DVA loss was 0.75 LogMAR, range = 0.44–1.23 LogMAR. This is significantly higher than has been previously documented in patients with bilateral vestibular hypofunction.[9]

Discussion

In this study, we evaluated a range of functional outcome measures in adults with NF2. Results demonstrated the inter- and intra-rater reliability of all chosen outcome measures were strong.

M9HPT. The m9HPT had excellent inter- and intra-rater reliability (ICC > 0.998), substantially higher than reliability for the same test in adults with NF1.[16] Rater reliability was also high for the classic nine-hole peg test in healthy adults, people with multiple sclerosis[23] and myotonic dystrophy type 1.[24] Mean score for m9HPT in NF2 was similar to scores in NF1[16] with mean 19.50 s (10.19–51.04) for the left hand, 18.55 s (10.29–40.09) for the right, but due to greater rater reliability, SEM and MDC were far better in NF2.

FSST. Inter- and intra-rater reliability of the FSST were excellent (ICC > 0.99) with tight confidence intervals. This is comparable to rater reliability for similar patient populations including older adults with a falls history[13] and people with multiple sclerosis.[25] SEM and MDC were small at 0.21 and 0.58 s, respectively. This is smaller than in multiple sclerosis (SEM = 1.67 s, MDC = 4.6 s),[25] which may reflect greater clinical homogeneity in NF2 than multiple sclerosis. The FSST correlated significantly with the subjective NFTI-QOL questionnaire question 8 (walking) but interestingly, not for question 1 which refers to balance, which may be due to how participants conceptualise balance. In this sample of 29 adults with NF2, the mean FSST score was 10.05 s (6.13–17.72). This is slightly faster than adults with multiple sclerosis (24) and vestibular dysfunction in a non-NF2

population.[26] Non-NF2 adults with a falls history and vestibular dysfunction, under the age of 65 years, scored mean 12.4 s ±4.2 SD and over 65 years, mean scored slower at 14.8 s (±4.3 s SD). In the study evaluating FSST in multiple sclerosis, researchers were able to differentiate 'fallers' from 'non-fallers' with FSST scores of 13.9 ± 6.9 SD in 'fallers' and 8.4 ± 1.8 SD 'non-fallers' (24). In adults over 65 years, Dite and Temple[13] found that 'fallers' scored a mean 23.59 s on the FSST, whereas 'non-fallers' scored 12.01 s. The cohort in this study were subcategorised by self-reported falls history and as previously outlined, although there was a trend towards worsening FSST score with a falls history, it did not reach statistical significance. This may be due to inequality in the sizes of the small subcategory comparison groups and would benefit from being revisited in a larger study, powered for this as the primary outcome. Whitney et al.[26] hypothesised that the FSST was not very sensitive for people with vestibular dysfunction as theoretically, one could keep the head still while performing the test.

Modified clinical test of sensory integration and balance. Inter- and intra-rater reliability were also excellent for the mCTSIB (ICC > 0.97), with tight confidence intervals. It was more difficult to compare reliability markers for the mCTSIB due to differences in testing protocols between studies meaning that comparisons were not possible. This highlights the need for standardised testing protocols. Adults attending an outpatient balance clinic had ICC 0.53–0.81,[27] but unlike this study, they completed the testing on a force plate (an instrumented mCTSIB) and classified completion of each subsection as maintaining balance for 10 s, rather than 30 s. Postural sway was computed through the force plate. We chose to complete the non-instrumented version of mCTSIB in this study so that it could be used by clinical team members in a variety of clinical locations, not restricted by access to force plates. In this study, SEM was 3.48 s in the mCTSIB composite score, with 9.66 s MDC. This was deemed reasonable for a test with mean score of 87.35 s. The mCTSIB test correlated with the NFTI-QOL questionnaire total score and subsections for walking and balance with the strongest correlation with the balance question. mCTSIB scores also significantly correlated with participants self-reported falls history. A total sum score of less than 88 s differentiated 'fallers' from 'non-fallers'. This is a finding with significant clinical implications and deserves further exploration to ascertain whether mCTSIB score can predict falls in NF2, and clearly identify when interventions such as balance rehabilitation should be initiated. The mCTSIB test had closer alignment with a patient's perceived balance experience than the FSST. In this study, the mean mCTSIB score was 87.35 s (35.8–120). To the best of our knowledge, mean sum scores for the mCTSIB have not been published in other populations/reliability studies, and published data for this test focuses on postural sway data obtained from force footplates[28,29] or completion of each subsection of the test for the allocated 10/30 s[27,30] meaning that we are unable to compare our findings with other studies. SEM and MDC of 3.48 and 9.66 s, respectively, for the mCTSIB sum score in this sample, were deemed appropriate.

All participants were able to maintain independent static standing with feet together on a stable surface for the full 30 s duration in condition 1 mCTSIB. There was greater heterogeneity in scores for subsections 2, 3 and 4, and 3/29 participants scoring maximum marks (120) for all four items, indicating that there may be a ceiling effect for the test. Interestingly, 2/29 participants, who obtained full scores in mCTSIB, had not fallen over but reported near misses in the past 12 months. FSST results for these participants were greater than 12 s, above the 10.05 s mean in this study, indicating that both tests may have a valuable place in evaluating balance deficits in this patient population.

DVA. Anecdotally, people with NF2 often describe difficulties viewing a stable target while their head is moving, for example, reading a train notice board while walking towards it. Findings from this study indicate that people with NF2 have significantly impaired VOR function and markedly reduced gaze stability in comparison to other patients with bilateral vestibular hypofunction.[9] Such large losses in DVA may be related to the high tumour burden of our population, although further investigation would be warranted in a larger sample. Vestibular rehabilitation has been shown to be effective in both patients with bilateral vestibular hypofunction,[9] and more specifically NF2.[31] Using DVA testing as an objective marker of functional VOR recovery in NF2 patients may be an avenue worth exploring in future work.

Long-term use of outcome measures. Functional outcome measures are used routinely to evaluate task performance over time and in response to treatment. The outcome measures should be clearly defined, specified in detail and measured consistently so that meaningful comparisons can be made.[32] Rater reliability for relatively objective measures such as timed performance measures, should be high, and were for all measures in this study. The measures were easy to conduct in the clinical outpatient setting, took less than a total of 5 min to complete all tests and were acceptable by both patients and the range of healthcare professionals who evaluated them, meaning that it would be feasible to continue to use them in practice. The mean mCTSIB sum scores aligned significantly with subjective experience of balance as rated in the NFTI-QOL questionnaire.

Limitations of the study. The sample size for this study was initially intended to meet the objective of ascertaining inter- and intra-rater reliability in NF2. We aimed to recruit 50 participants but achieved 29. Twenty-nine participants were sufficient to confidently evaluate rater reliability in this study as reliability markers were greater than 0.9 with tight confidence intervals for all measures.[33] This does lead to

questions about whether outcome measurement findings in this study (mean values, SD, range, etc.) are representative of the wider NF2 population internationally and this would benefit from further investigation in a larger multisite study, with a sample size calculation powered for this.

Conclusion

There is a need for reliable functional outcome measures to evaluate disease progression and monitor treatment in NF2. The mCTSIB, FSST and m9HPT had excellent rater reliability and were acceptable to both patients and professionals using them in the clinic setting. The mCTSIB was also able to discriminate fallers from non-fallers. We intend to evaluate the outcomes for test–retest reliability and validity in multi-centre and longitudinal studies.

Acknowledgements

The authors acknowledge Fiona Reid and Alexandra Curtis for their valuable contributions to this study.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Ethical approval

Ethical approval was granted for this study by National Research Ethics Service Committee North West, reference 16/NW/0504.

Funding

The author(s) received no financial support for the research, authorship and/or publication of this article.

Informed consent

Written informed consent was obtained from all participants in this study, including age 16–18 years. In the United Kingdom, young people over the age of 16 years are deemed able to provide informed consent when 'Gillick competent'. This includes all studies that do not involve clinical trials of an interventional product (CTIMPS) (http:///wwww.hradecisiontools.org.uk/consent/principles-children-EngWalesNI.html).

Trial registration

This trial is registered at clinicaltrials.gov identifier: NCT03617276.

ORCID iD

Rebecca Louise Mullin, https://orcid.org/0000-0002-2571-104X

Supplemental material

Supplemental material for this article is available online.

References

1. Lloyd SKW and Evans DGR. Neurofibromatosis type 2 (NF2). In: Said G and Krarup C (eds) *Handbook of clinical neurology, peripheral nerve disorders*, vol. 115. Edinburgh: Elsevier, 2013, pp. 957–967.
2. Ferner RE, Shaw A, Evans DG, et al. Longitudinal evaluation of quality of life in 288 patients with neurofibromatosis 2. *J Neurol* 2014; 261(5): 963–969.
3. Morris KA, Golding JF, Axon PR, et al. Bevacizumab in neurofibromatosis type 2 (NF2) related vestibular schwannomas: a nationally coordinated approach to delivery and prospective evaluation. *Neurooncol Pract* 2016; 3(4): 281–289.
4. de Vet HCW, Terwee CB, Mokkink LB, et al. *Measurement in medicine*. Cambridge: Cambridge University Press, 2011.
5. Mokkink LB, Terwee CB, Patrick DL, et al. The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. *J Clin Epidemiol* 2010; 63(7): 737–745.
6. Stokes E. *Rehabilitation outcome measures*. Edinburgh: Churchill Livingstone, 2011.
7. Stratford PW, Binkley JM and Riddle DL. Health status measures: strategies and analytic methods for assessing change scores. *Phys Ther* 1996; 76(10): 1109–1123.
8. Hermann R, Ionescu EC, Dumas O, et al. Bilateral vestibulopathy: vestibular function, dynamic visual acuity and functional impact. *Front Neurol* 2018; 9: 555.
9. Herdman SJ, Hall CD, Schubert MC, et al. Recovery of dynamic visual acuity in bilateral vestibular hypofunction. *Arch Otolaryngol Head Neck Surg* 2007; 133(4): 383–389.
10. Herdman SJ, Tusa RJ, Blatt P, et al. Computerized dynamic visual acuity test in the assessment of vestibular deficits. *Am J Otol* 1998; 19(6): 790–796.
11. Dannenbaum E, Paquet N, Chilingaryan G, et al. Clinical evaluation of dynamic visual acuity in subjects with unilateral vestibular hypofunction. *Otol Neurotol* 2009; 30(3): 368–372.
12. Gargon E, Gurung B, Medley N, et al. Choosing important health outcomes for comparative effectiveness research: a systematic review. *PLoS ONE* 2014; 9: e99111.
13. Dite W and Temple VA. A clinical test of stepping and change of direction to identify multiple falling older adults. *Arch Phys Med Rehabil* 2002; 83(11): 1566–1571.
14. Moore M and Barker K. The validity and reliability of the four square step test in different adult populations: a systematic review. *Syst Rev* 2017; 6: 187.
15. Cohen H, Blatchly CA and Gombash LL. A study of the clinical test of sensory interaction and balance. *Phys Ther* 1993; 73(6): 346–351; discussion 351.
16. Mullin RL, Golding JF, Smith R, et al. Reliability of functional outcome measures in adults with neurofibromatosis 1. *SAGE Open Med* 2018; 6: 2050312118786860.
17. Strupp M, Kim JS, Murofushi T, et al. Bilateral vestibulopathy: diagnostic criteria consensus document of the Classification Committee of the Bárány Society. *J Vestib Res* 2017; 27(4): 177–189.
18. Hornigold RE, Golding JF, Leschziner G, et al. The NFTI-QOL: a disease-specific quality of life questionnaire for neurofibromatosis 2. *J Neurol Surg B Skull Base* 2012; 73(2): 104–111.
19. Botolfsen P, Helbostad JL, Moe-Nilssen R, et al. Reliability and concurrent validity of the expanded timed up-and-go test in older people with impaired mobility. *Physiother Res Int* 2008; 13(2): 94–106.

20. Flansbjer UB, Holmbäck AM, Downham D, et al. Reliability of gait performance tests in men and women with hemiparesis after stroke. *J Rehabil Med* 2005; 37(2): 75–82.
21. Poncumhak P, Saengsuwan J, Kamruecha W, et al. Reliability and validity of three functional tests in ambulatory patients with spinal cord injury. *Spinal Cord* 2013; 51(3): 214–217.
22. Koo TK and Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *J Chiropr Med* 2016; 15: 155–163.
23. Cohen JA, Fischer JS, Bolibrush DM, et al. Intrarater and interrater reliability of the MS functional composite outcome measure. *Neurol* 2000; 54: 802–806.
24. Cutellè C, Rastelli E, Gibellini M, et al. Validation of the Nine Hole Peg Test as a measure of dexterity in myotonic dystrophy type 1. *Neuromuscul Disord* 2018; 28(11): 947–951.
25. Wagner JM, Norris RA, Van Dillen LR, et al. Four Square Step Test in ambulant persons with multiple sclerosis. *Int J Rehabil Res* 2013; 36(3): 253–259.
26. Whitney SL, Marchetti GF, Morris LO, et al. The reliability and validity of the four square step test for people with balance deficits secondary to a vestibular disorder. *Arch Phys Med Rehabil* 2007; 88(1): 99–104.
27. Loughran S, Tennant N, Kishore A, et al. Interobserver reliability in evaluating postural stability between clinicians and posturography. *Clin Otolaryngol* 2005; 30(3): 255–257.
28. Suttanon P, Hill KD, Dodd KJ, et al. Retest reliability of balance and mobility measurements in people with mild to moderate Alzheimer's disease. *Int Psychogeriatr* 2011; 23(7): 1152–1159.
29. Klima D, Morgan L, Baylor M, et al. Physical performance and fall risk in persons with traumatic brain injury. *Percept Mot Skills* 2019; 126(1): 50–69.
30. Deshpande N, Hewston P and Aldred A. Sensory functions, balance, and mobility in older adults with type 2 diabetes without overt diabetic peripheral neuropathy: a brief report. *J Appl Gerontol* 2017; 36(8): 1032–1044.
31. Emmanouil B, Browne K, Halliday D, et al. First report of the efficacy of vestibular rehabilitation in improving function in patients with Neurofibromatosis type 2: an observational cohort study in a clinical setting. *Disabil Rehabil* 2019; 41(14): 1632–1638.
32. Smith P, Morrow R and Ross D. *Field trials of health interventions: a toolbox*. 3rd ed. Oxford: Oxford University Press, 2015.
33. Donner A and Eliasziw M. Sample size requirements for reliability studies. *Stat Med* 1987; 6: 441–448.