



## Reliability of the North Star Ambulatory Assessment in a multicentric setting

E.S. Mazzone<sup>a,1</sup>, S. Messina<sup>a,b,1</sup>, G. Vasco<sup>a</sup>, M. Main<sup>c</sup>, M. Eagle<sup>d</sup>, A. D'Amico<sup>e</sup>, L. Doglio<sup>f</sup>, L. Politano<sup>g</sup>, F. Cavallaro<sup>b</sup>, S. Frosini<sup>h</sup>, L. Bello<sup>i</sup>, F. Magri<sup>j</sup>, A. Corlatti<sup>k</sup>, E. Zucchini<sup>l</sup>, B. Brancalion<sup>m</sup>, F. Rossi<sup>n</sup>, M. Ferretti<sup>o</sup>, M.G. Motta<sup>o</sup>, M.R. Cecio<sup>g</sup>, A. Berardinelli<sup>o</sup>, P. Alfieri<sup>a,e</sup>, T. Mongini<sup>n</sup>, A. Pini<sup>l</sup>, G. Astrea<sup>h</sup>, R. Battini<sup>h</sup>, G. Comi<sup>j</sup>, E. Pegoraro<sup>i</sup>, L. Morandi<sup>k</sup>, M. Pane<sup>a</sup>, C. Angelini<sup>i</sup>, C. Bruno<sup>f</sup>, M. Villanova<sup>m</sup>, G. Vita<sup>b</sup>, M.A. Donati<sup>p</sup>, E. Bertini<sup>e</sup>, E. Mercuri<sup>a,c,\*</sup>

<sup>a</sup> Department of Paediatric Neurology, Catholic University, Rome, Italy

<sup>b</sup> Department of Neurosciences, Psychiatry and Anaesthesiology, University of Messina, Messina, Italy

<sup>c</sup> Dubowitz Neuromuscular Unit, Institute of Child Health, London, UK

<sup>d</sup> Institute of Human Genetics, University of Newcastle upon Tyne, UK

<sup>e</sup> Department of Laboratory Medicine, Unit of Molecular Medicine and Department of Pediatric Neuropsychiatry, Bambino Gesù Hospital, Rome, Italy

<sup>f</sup> Neuromuscular Disease Unit, G. Gaslini Institute, Genoa, Italy

<sup>g</sup> Dipartimento Medico Chirurgico di Internistica Clinica e Sperimentale, Seconda Università di Napoli, Italy

<sup>h</sup> Department of Paediatric Neurology, Stella Maris Institute, University of Pisa, Italy

<sup>i</sup> Department of Neurosciences, University of Padua, Padua, Italy

<sup>j</sup> IRCCS Foundation Ospedale Maggiore Policlinico Mangiagalli and Regina Elena, Milan, Italy

<sup>k</sup> Myopathology and Neuroimmunology, Pediatric Neurology, Neurological Institute C. Besta, Milan, Italy

<sup>l</sup> Child Neurology and Psychiatry Unit, Maggiore Hospital, Bologna, Italy

<sup>m</sup> Nigrisoli Hospital, Bologna, Italy

<sup>n</sup> Neuromuscular Center, S.G. Battista Hospital, University of Turin, Italy

<sup>o</sup> IRCCS "C. Mondino" Foundation, University of Pavia, Italy

<sup>p</sup> Metabolic and Neuromuscular Unit, Meyer Hospital, Florence, Italy

### ARTICLE INFO

#### Article history:

Received 6 March 2009

Received in revised form 5 May 2009

Accepted 4 June 2009

#### Keywords:

Duchenne Muscular Dystrophy

Outcome

Function

### ABSTRACT

The aim of this study was to investigate the suitability of the North Star Ambulatory Assessment as a possible outcome measure in multicentric clinical trials. More specifically we wished to investigate the level of training needed for achieving a good interobserver reliability in a multicentric setting.

The scale was specifically designed for ambulant children with Duchenne Muscular Dystrophy and includes 17 items that are relevant for this cohort. Thirteen Italian centers participated in the study. In the first phase of the study we provided two training videos and an example of the scale performed on a child. After the first session of training, all the 13 examiners were asked to send a video with an assessment performed in their centre and to score all the videos collected. There were no difficulties in performing the items and in obtaining adequate videos with a hand held camera but the results showed a poor interobserver reliability (<.5). After a second training session with review and discussion of the videos previously scored, the same examiners were asked to score three new videos. The results of this session had an excellent interobserver reliability (.995).

The level of agreement was maintained even when the same videos were rescored after a month, showing a significant intra-observer reliability (.95).

Our results suggest that the NSAA is a test that can be easily performed, completed in 10 min and can be used in a multicentric setting, providing that adequate training is administered.

© 2009 Published by Elsevier B.V.

### 1. Introduction

Steroids have been considered for over a decade the best therapeutic option to improve function or at least to slower clinical deterioration in Duchenne Muscular Dystrophy (DMD) [1]. In the

last few years however several other therapeutic approaches, such as the use of stem cells or antisense oligonucleotides, have become or are becoming available for phase 2/3 clinical trials. The planning of these trials has highlighted that the existing outcome measures are often not appropriate to provide accurate quantification of the effects of such therapies and to satisfy the needs of researchers, clinicians and regulatory agencies.

As part of TREAT NMD, a European network of excellence aimed at facilitating translational research in neuromuscular disorders,

\* Corresponding author. Address: Department of Child Neurology, Policlinico Gemelli, Largo Gemelli 00168, Roma, Italy. Tel.: +39 06 30155340; fax: +39 06 30154363.

E-mail address: [mercuri@rm.unicatt.it](mailto:mercuri@rm.unicatt.it) (E. Mercuri).

<sup>1</sup> Both authors contributed equally.

the issue of outcome measures in DMD has been recently systematically reviewed [2]. With a few exceptions, the measures used so far in DMD are measures of muscle strength, obtained using clinical assessments such as the Medical Research Council Scale (MRC), and Manual Muscle Testing (MMT) [3–7], or other structured assessments (hand held dynamometers and quantitative muscle testing) [6,8,9].

The assessment of strength however does not always reflect the subject's functional ability [2,10] and, as also suggested by FDA and other regulatory authorities, there is the need for measures that are 'clinically meaningful' to parents and families. Timed items, such as walking 10 m or the 6 min walk test (6MWT), have been increasingly used as they can reliably detect changes over time and, for the 10 m test, also predict age of loss of ambulation in DMD boys [11].

A few attempts have been made to identify other functional measures reflecting other aspects of everyday life activities. Classically, the most commonly used functional scales in DMD are those developed by Vignos et al. [12] and Brooke et al. [5] and the Hammersmith motor ability scale (HMAS) [3], largely used in clinical practice but for which there is little published information regarding its reliability and validity. In the last few years new scales have been developed to address some of these issues. Some of these, such as the Egen Klassifikation (EK) scale [13–15] the Motor Function Measure (MFM) [10,16] provide a mean of assessment that covers the whole range of activities from very weak non ambulant patients to strong ambulant patients and have also been validated in subgroups of patients affected by DMD.

The North Star Ambulatory Assessment (NSAA) is a new functional scale specifically designed for ambulant DMD boys trying to address some of the shortcomings reported for the other scales. The NSAA has been based on the HMAS, revising most of the items from the previous version that were felt to be reliable indicators of possible functional changes in ambulant DMD boys but also includes a number of new items that provide the opportunity to detect possible improvement following treatment, avoiding the ceiling effect criticised in the other existing scales. Some of these activities, such as head raise, hopping and running are usually not observed in untreated DMD children but are increasingly seen in children treated with daily steroids and should be assessed and monitored following new treatments. The scale has been piloted by the North Star Clinical Network for Paediatric Neuromuscular Disease Management (NSCN) in the UK in the United Kingdom with good intra and interobserver reliability [2,17,18] and has recently been suggested to be used as one of the outcome measures in forthcoming studies.

The aim of this study was to further assess inter and intra-observer reliability of the NSAA and the level of training needed in a multicentric setting.

## 2. Methods and subjects

### 2.1. NSAA

The scale consists of 17 items (Table 1), ranging from standing (item 1) to running (item 17) and includes several items assessing abilities that are necessary to remain functionally ambulant i.e. ability to rise from the floor, ability to get from lying to sitting and sitting to standing and that are known to progressively deteriorate in untreated DMD patients. The scale also includes items assessing head raise and standing on heels that can be partly present in the early stages of the disease and a number of activities such as hopping, jumping and running that are generally never fully achieved in untreated DMD boys but that have been found in those treated with daily steroids.

**Table 1**

Percentage of agreement on the scores of individual items and total scores in phases 1 and 2.

	Test item	Phase 1	Phase 2
1.	Stand	96.5	100
2.	Walk (10 m)	95	97
3.	Sit to stand from chair	97.7	100
4.	Stand on one leg – R	77.7	100
5.	Stand on one leg – L	78	100
6.	Climb step – R	97	100
7.	Climb step – L	98.2	100
8.	Descend step – R	96.5	100
9.	Descend step – L	89.2	100
10.	Gets to sitting	84.5	97
11.	Rise from floor	91.6	100
12.	Lifts head	74.3	100
13.	Stand on heels	71.8	100
14.	Jump	91.5	100
15.	Hop – R	68.5	92
16.	Hop – L	80	100
17.	Run	78.5	92
	Mean agreement	87	98.7
	No. of items above 90%	8/17	17/17
	Total scores	33	90%

Instructions on how to elicit the items are available on the proforma but more detailed instructions and additional information on how to score individual items is available and can be downloaded at the TREAT NMD site (<http://www.researchchrom.com/masterlist/view/18>).

Each item can be scored on a three point scale using simple criteria: 2 – 'Normal' – achieves goal without any assistance; 1 – Modified method but achieves goal independent of physical assistance from another; 0 – Unable to achieve independently.

A total score can be achieved by summing the scores for all the individual items. The score can range from 0, if all the activities are failed, to 34, if all the activities are achieved. All items have to be tested without thoracic braces or leg orthoses. The scale is generally completed in a maximum of 15 min.

The scale also includes the possibility to record timed items (walk 10 m and rise from the floor). The time taken to complete the task is not part of the global score but provides an additional measure of the DMD boys'abilities that can be monitored over time.

### 2.2. Training sessions

One of the Italian physiotherapists involved in the study (ESM) was trained in London and in Newcastle by two experienced physiotherapists (MM and ME) who had been involved in the development of the NSAA and in the UK training program for NSAA. The training consisted in assessing several patients together, having the opportunity to discuss results and possible bottlenecks in performing and scoring, followed by scoring of training videos.

### 2.3. Dissemination of the training in the Italian centers

The physiotherapist trained in UK organised training sessions with physiotherapists from each of the 13 Italian tertiary care pediatric neuromuscular centers. Three training sessions were organised in order to have a limited number of people attending each session. As one of the aims of this study was to establish the level of training needed to have an adequate interobserver agreement, the first training session (phase 1) mainly consisted in a presentation of the scale, two training videos and the opportunity to have a demonstration on a patient to see how the scale has to be performed. All the therapists attending the training

sessions had experience with neuromuscular disorders and with functional scales in DMD but none had used the NSAA before the training session.

After these sessions the examiners from the 13 centers were asked to perform a video in one of their patients after they had been provided of instructions on how to perform the video and the manual with instructions on scoring. The videos were collected on a CD and circulated among the participants to get their scores. The aim of this part was to establish the quality of the videos, whether the assessments had been performed correctly, and the ability to score.

The children assessed in the 13 centers were all ambulant DMD patients with age ranging between 5 years and 9 years, 11 months. All but two were on steroids and their level of function, assessed by NSAA ranged between 17 and 32/34 (mean 24, 81).

#### 2.4. Phase 2

A second session (phase 2) including all the participants to the study was organised after analysis of the videos in order to look at the consistency of the results. In this session all participants were asked to look at the videos previously scored and discuss possible differences in performing and scoring the individual items. All the videos were shown in 'real time' without the opportunity to rewind and look at the clips more than one time. After this second training session the participants were asked to score three new videos.

Intrarater reliability was determined by a repeated evaluation of two of the videos scored in phase 2 one month later.

#### 2.5. Statistical analysis

Statistical Package for the Social Science 11.0.1 for Windows (SPSS, Chicago, IL) was used for data analysis (reliability, construct validity and internal consistency) [19,20]. To assess rating reliability the Intraclass Correlation Coefficient (ICC) was calculated by comparing the variability of different ratings of the same subject to the total variation across all ratings and all subjects. The Intraclass Correlation Coefficient typically range from 0 to 1 with scores closer to 1 representing greater reliability.

Other than having a coefficient of reliability, we were also interested to look at the consistency of the scores of the individual items in order to identify individual items for which agreement may be more difficult to reach. The results of the individual items were compared to a standard which was set at 90% i.e. it was expected that for any one test item the group would be in agreement as to scoring 9 times out of 10, as a practical criterion.

The test–retest reliability was assessed with the Spearman–Brown coefficient to estimate full test reliability based on split-half reliability measures. This statistical analysis measures the equivalence of the scores during the evaluation of the same items with the same instrument on the same subjects. Test–retest reliability with the split-half method allows to assess internal consistency.

### 3. Results

After the first phase of training, all the centers sent their video, as requested. Two of the three were in a format that did not allow easy transfer to a CD. The remaining 11 were sent on a CD and were scored by all centers. The quality of the video was acceptable in all but in four of the 11 recordings there were two items in which scoring was not possible (e.g. feet could not be visualised when jumping or videos taken from the wrong

angle when going on stairs), for a total of 179 out of 187 scorable items (95.7%).

#### 3.1. Score agreement

In phase 1 there was poor agreement among the examiners. For each video the percentage of similar total scores ranged from 23 to 53% (mean 33%) that increased to 38 to 84% (mean 42.2%) when a possible difference of +1 point was considered. The ICC was very poor (<.5).

When individual items were analyzed, the level of agreement among the different examiners ranged from 68.5 to 98.2, and 9 of the 17 items did not meet the standard (90%).

The items that more often had less satisfactory results in the first phase and that required further training were:

*Standing on one leg (items 4 and 5)*: there was no agreement on the difference between 'no fixation' (score 2) and 'needs a lot of fixation' score 1, with difficulties in scoring children who had only 'a little fixation' but not 'a lot'. In agreement with the instructions provided by the therapists who developed the method, it was clarified that 'a little fixation' is allowed and can be scored as 'no fixation' (score 2).

*Lifts head (item 12)*: some examiners scored as 2 any attempt to lift head from the plinth irrespective of the method used, even if the instructions on the manual provided detailed description of how to score it.

*Stands on heels (item 13)*: some examiners tended to score 0 all the patients who had eversion of the foot even when this was associated with obvious forefoot lifting (score 1).

*Hopping (item 15 and 16)*: some examiners scored as 2 even in patients who got heel only off the ground and not forefoot.

*Running (item 17)*: some of the examiners felt that a good Duchenne jog should be scored as 2 as it was better than a poor one that was still scored as 1.

In phase 2, after the second session of training there was a very good agreement of the global scores with all examiners giving the same final score to two of the three videos while in the third video 9/13 examiners gave the same score and the other four gave a score different of one point. The ICC calculated on the three global scores of the three videos was of .995. When individual items were analysed the ICC ranged between .94 and 1 in 16 of the 17 items and was .75 in one item (item 15). All 17 items were found to meet the standard (90%). Table 1 provides details of the agreement on individual items.

The test–retest reliability measured for all examiners by Spearman–Brown split-half reliability ranged from .70 to 1 for B2 and from .72 to 1 for B3.

### 4. Discussion

In this paper we report the use of the NSAA, a simple inexpensive tool largely used in clinical settings in order to establish its suitability in research trials. One of the advantages of the NSAA is that it was specifically designed for ambulant children with Duchenne Muscular Dystrophy and only includes items that are relevant for this cohort.

Unlike other more general scales for neuromuscular disorders, that also includes items assessing other aspects of function relevant to other neuromuscular disorders or to more advanced phases of DMD, in the NSAA the number of items is relatively small. The scale takes approximately 10–15 min and can be easily performed even in children with moderate learning difficulties or some behavioral problems. Another advantage is that it also includes timed tests that are routinely used in many centers to monitor clinical changes. This avoids repetition which may induce

fatigue or lack of co-operation and leaves time for other assessments such as the 6MWT that are becoming increasingly used in both clinical and research settings.

Reliability and feasibility of the scale have already been reported in a UK study that has also reported how the NSAA can detect changes over time in DMD boys [17,18]. In the present study we have further explored the reliability of the scale in order to establish the suitability of the scale for multicentric studies.

In the first phase of the study we were interested to establish whether a simple training showing two videos and an example of the scale performed on a child in front of a small number of trainees could be enough to achieve good interobserver agreement among examiners who had experience with other neuromuscular functional scales but who had never performed the NSAA before the study. The analysis of the results showed a poor agreement on global scores for 9 of the 17 items. It is of interest that all the 13 examiners were able to perform the scale eliciting each item correctly but the interpretation of the results was dubious and further discussion with other examples was needed to find an acceptable agreement. While there were little doubts on the items failed (score 0), there were some discrepancies between the scores 1 and 2 and on the interpretation of the general guidelines provided in the manual. Even if most of the responses were available on the manual, some examiners needed an extra session to clarify some of these issues, especially when the items were similar but were scored differently in other scales with which they were more familiar.

The review of the videos was very useful and an agreement was found on what should be interpreted as a full score. Three new videos scored after the second phase of training showed much better results with values of agreement ranging between 92 and 100%.

The level of agreement was maintained when the same videos were rescored after a month, showing significant intra-observer reliability.

In conclusion, our results suggest that the NSAA is a practical test that can be easily completed in 10 min and can be used in a multicentric setting, as long as adequate training is provided. There were no difficulties in performing the items and in obtaining adequate videos with a hand held camera, even after a short training session. In contrast, there was the need for a second session to have a better interobserver reliability when scoring. Further studies are in progress to explore the ability of NSAA to detect changes over time in relation to other measures of function such as 6MWT and myometry.

## Acknowledgments

The study is supported by a Telethon UILDM Grant (GUP07009) and by TREAT NMD. The authors thank Giuseppina Sgandurra, for statistical advice and Francesco Muntoni, Kate Bushby, Adnan Manzur, Elaine Scott and Meredith Jones for their help and useful suggestions.

## References

- [1] Manzur AY, Kuntzer T, Pike M, Swan A. Glucocorticoid corticosteroids for Duchenne muscular dystrophy. *Cochrane Database Syst Rev* 2008;CD003725.
- [2] Mercuri E, Mayhew A, Muntoni F, Messina S, Straub V, Van Ommen GJ, et al. On behalf of the TREAT-NMD Neuromuscular Network. Towards harmonisation of outcome measures for DMD and SMA within TREAT-NMD; Report of three expert workshops: TREAT-NMD/ENMC workshop on outcome measures, 12th–13th May 2007, Naarden, The Netherlands; TREAT-NMD workshop on outcome measures in experimental trials for DMD, 30th June–1st July 2007, Naarden, The Netherlands; Conjoint Institute of Myology TREAT-NMD Meeting on physical activity monitoring in neuromuscular disorders, 11th July 2007, Paris, France. *Neuromuscul Disord* 2008;18:894–903.
- [3] Scott OM, Hyde SA, Goddard C, Dubowitz V. Quantisation of muscle function in children: a prospective study in Duchenne muscular dystrophy. *Muscle Nerve* 1982;5:291–301.
- [4] Florence JM, Pandya S, King WM, et al. Intrarater reliability of manual muscle test (Medical Research Council scale) grades in Duchenne's muscular dystrophy. *Phys Ther* 1992;72:115–22.
- [5] Brooke MH, Griggs RC, Mendell JR, Fenichel GM, Shumate JB, Pellegrino RJ. Clinical trial in Duchenne dystrophy. I. The design of the protocol. *Muscle Nerve* 1981;4:186–97.
- [6] Escolar DM, Henricson EK, Mayhew J, et al. Clinical evaluator reliability for quantitative and manual muscle testing measures of strength in children. *Muscle Nerve* 2001;24:787–93.
- [7] Kilmer DD, Abresch RT, Fowler Jr WM. Serial manual muscle testing in Duchenne muscular dystrophy. *Arch Phys Med Rehabil* 1993;74:1168–71.
- [8] Stuberg WA, Metcalf WK. Reliability of quantitative muscle testing in healthy children and in children with Duchenne muscular dystrophy using a hand-held dynamometer. *Phys Ther* 1988;68:977–82.
- [9] Mayhew JE, Florence JM, Mayhew TP, et al. Reliable surrogate outcome measures in multicenter clinical trials of Duchenne muscular dystrophy. *Muscle Nerve* 2007;35:36–42.
- [10] Bérard C, Payan C, Hodgkinson I, Fermanian J, and the MFM Collaborative Study Group. A motor function measure scale for neuromuscular diseases. Construction and validation study. *Neuromuscul Disord* 2005;15:463–70.
- [11] McDonald CM, Abresch RT, Carter GT, et al. Profiles of neuromuscular diseases. Duchenne muscular dystrophy. *Am J Phys Med Rehabil* 1995;74:S70–92.
- [12] Vignos Jr PJ, Spencer Jr GE, Archibald KC. Management of progressive muscular dystrophy in childhood. *JAMA* 1963;13:89–96.
- [13] Steffensen B, Hyde S, Lyager S, Mattsson E. Validity of the EK scale: a functional assessment of non-ambulatory individuals with Duchenne muscular dystrophy or spinal muscular atrophy. *Physiother Res Int* 2001;6:119–34.
- [14] Steffensen BF, Hyde SA, Attermann J, Mattsson E. Reliability of the EK scale, a functional test for non-ambulatory persons with Duchenne dystrophy. *Adv Physiother* 2002;4:47.
- [15] Steffensen BF, Lyager S, Werge B, Rahbek J, Mattsson E. Physical capacity in non-ambulatory people with Duchenne muscular dystrophy or spinal muscular atrophy: a longitudinal study. *Dev Med Child Neurol* 2002;44:623–32.
- [16] Bérard C, Payan C, Fermanian J, Girardot F, Groupe d'Etude MFM. A motor function measurement scale for neuromuscular diseases – description and validation study. *Rev Neurol (Paris)* 2006;162:485–93. [In French].
- [17] Scott E, Eagle M, Main M, Sheehan J. The North Star Ambulatory Assessment. Abstract 31st annual meeting of the British Paediatric Neurology Association. 18th–20th January 2006. *Dev Med Child Neurol* 2006;2006:27.
- [18] Eagle M, Scott E, Main M, Sheehan J, Michelle M, Guglieri M, et al. Steroids in Duchenne muscular dystrophy (DMD): natural history and clinical evaluation using the North Star Ambulatory Assessment (NSAA). Abstract World Muscle Society, Taormina, Italy 17–20 October 2007, *Neuromuscul Disord* 2007;17:774.
- [19] MacLennan RN. Interrater reliability with SPSS for Windows 5.0. *Am Stat* 1993;47:292–6.
- [20] Portney LG, Watkins MP. Foundation of clinical research applications to practise. Norwalk, CT: Appleton and Lange; 1993.