



PERGAMON

Neuromuscular Disorders xx (2006) 1–10



www.elsevier.com/locate/nmd

A modified Hammersmith functional motor scale for use in multi-center research on spinal muscular atrophy

Kristin J. Krosschell ^{a,*}, Jo Anne Maczulski ^b, Thomas O. Crawford ^c,
Charles Scott ^d, Kathryn J. Swoboda ^e

^a Department of Physical Therapy and Human Movement Sciences, Feinberg School of Medicine, Northwestern University, Suite 1100, 645 North Michigan Avenue, Chicago, IL 60611, USA

^b 537 West Roscoe, Chicago, IL, USA

^c Johns Hopkins Departments of Neurology and Pediatrics, Johns Hopkins University, Baltimore, MD, USA

^d Biostatistics and Data Management Core, Childrens Hospital of Philadelphia, Philadelphia, PA, USA

^e University of Utah School of Medicine, Salt Lake City, Utah

Received 19 December 2005; received in revised form 16 March 2006; accepted 29 March 2006

Abstract

The Hammersmith functional motor scale for children with spinal muscular atrophy was modified to establish a standard measure of functional ability in children with non-ambulant spinal muscular atrophy types 2 and 3 in a longitudinal multi-center clinical trial. This study assessed the intra- and interrater reliability and the test–retest stability of a modified version of the scale. Both intra- and interrater reliability were established. Results indicate that the scale is reliable and stable over a 6 month period. Reliability was maintained when patient sample criteria were expanded to include children younger than 30 months and children with popliteal angles greater than 20°. These data establish the modified Hammersmith functional motor scale for children with spinal muscular atrophy as a reliable instrument for use in multi-center treatment trials in non-ambulant spinal muscular atrophy children. Our data provides additional support for the use of original scale items in terms of ease of administration, usefulness and reliability, while incorporating modifications to optimize its use in a multi-center clinical research setting.

© 2006 Published by Elsevier B.V.

Keywords: Spinal muscular atrophy; Functional outcome measure; Clinical trials

1. Introduction

Recent investigations into the pathogenesis of spinal muscular atrophy (SMA) have led to hope that a specific therapy might soon be possible [1,2]. This prospect raises important concerns about the best manner for testing a putative therapy [3]. Critical to the success of any treatment trial is the development and validation of outcome measures that are (i) sensitive to small changes in functional ability, (ii) highly reliable in their application, (iii) appropriate for use in young children, and (iv) have acknowledged face

validity as a meaningful measure of the burden of the disease.

SMA manifests across a spectrum of severity, from those with severe weakness beginning in early infancy to those with mild weakness with onset in adulthood. Determining which subjects are likely to be best suited to demonstrate a treatment effect is of substantial import. Many factors influence this choice, including the quality of outcome measures for a given cohort, the number of potential research subjects in each cohort, and the presence of confounding factors. Easily identifiable subgroups for SMA treatment trials are currently four in number:

- Infants with severe SMA, though abundant, present special difficulties to experimental design because their young age makes measurement of function difficult and their extreme weakness makes them vulnerable to many

* Corresponding author. Tel.: +1 312 908 6790; fax: +1 847 256 8134.
E-mail address: k-krosschell@northwestern.edu (K.J. Krosschell).

complications that confound ascertainment of treatment effect.

- Adults and older children who have experienced advancing weakness for only a short duration are attractive subjects for therapeutic trials because the reliable outcome measures developed for ALS would be readily adaptable. Recruitment of sufficient patients may be difficult, however, because these patients are not abundant.
- Study of adults and older children with long standing weakness offers the advantage of highly motivated subjects capable of reliable outcome measurement, but the relative stability of their disorder raises other concerns. It is likely that confounding effects such as severe contractures, replacement of long-denervated muscle with fibrofatty tissue, lung disease, and other complications may limit responsiveness. In addition, the static clinical course suggests the possibility that SMN abundance no longer has a major role in disease progression.
- The remaining group: toddlers and younger children manifesting intermediate levels of weakness, offers certain relative advantages, but also some distinct challenges. At present they represent the most prevalent group, and their families are highly motivated to enroll in trials. The chief problem with focusing study upon this group, however, is the difficulty in outcome assessment. Development and validation of a quality outcome measure to assess treatment effect in children who are capable of only limited understanding of the trial, and hence limited ability to be fully motivated to cooperate, is of great import.

Currently, it is an unfortunate reality that earlier in the course of the disease when we might best be able to influence disease course, outcome measures are limited. However, at older ages when cooperation is better, there are fewer patients to be studied, and secondary complications of the disease are more likely to constrain measurement of benefit. Our goal in assessing and modifying the current Hammersmith instrument is to help address this important issue, making it more readily available for use in a research setting in young children with SMA.

In most neuromuscular disorders, direct measures of muscle power with myometry and other quantitative tests of strength are the most accepted, as there is immediate face validity [4,5]. Miller et al. [6], Merlini et al. [5,7,8] and others [9] have established the reliability and validity of myometry in the assessment of strength in children with SMA, but only in those older than age 5. On the other hand, functional motor scales may be more appropriate for young children, as motivation for maximum performance need not depend upon comprehension of the purpose of the task. Thus, scales targeted to assess functional ability with disease-specific and strength-specific tasks may have the potential to exceed direct measures of power in sensitivity and reliability

[10–12]. In addition, functional testing to ascertain efficacy and monitor natural history has value [13,14]. Functional tests, however, are often limited by lack of sensitivity and/or the inability of the test to efficaciously monitor change over the full course/spectrum of the disease. Functional testing may not be able to detect subtle changes and/or monitor for changes in muscle strength [15]. However, advantages may outweigh disadvantages in that functional tests are better at assessing outcomes (activities of daily living) that are more meaningfully appreciated by patients.

Functional measures developed specifically for older patients with neuromuscular diseases have included the Vignos lower extremity classification scale [16], the Brooke upper extremity scale [17,18], various timed functional assessments [17], and The Hammersmith motor ability scale [19]. More recently the EK scale [20] was developed for non-ambulatory patients with Duchenne's muscular dystrophy (DMD) and SMA, the Wee Fim was utilized to quantify function in children with SMA in Hong Kong [11] and the functional research scale for ALS (FRS-ALS) scale was utilized in clinical trials with adult patients with ALS [21,22]. These scales are typically used as primary outcome measures for a treatment trial only when direct measure of power is not possible, though their use as a secondary measure is common [5,6,23–26]. Though each of these tests were developed to assess functional skills in weak patients with neuromuscular disorders, construct of test items often precludes their use in assessing younger children.

The Hammersmith functional motor scale for children with SMA was recently developed after careful assessment of important functional skills of normal children that are compromised in non-ambulatory children with SMA [10] to provide a tool for effective clinical assessment of motor abilities in this population. The Hammersmith functional motor scale for children with SMA is intended to be sensitive to those functional motor deficits of children with SMA that result directly from weakness. In a single institution setting, the Hammersmith functional motor scale for children with SMA has been shown to be both quick to administer and reliable in non-ambulant children with SMA as young as 30 months of age [10], when other tools such as myometry or the medical research council (MRC) scale cannot be easily or reliably performed. Initially intended to provide a better means for assessment of functional level for individual children with SMA, we saw this scale as having potential value, with further modification, for multi-center and longitudinal use, as an outcome measure for treatment trials.

This study was intended to prepare and evaluate a modified version of the Hammersmith functional motor scale for children with SMA for use in a multi-center collaborative setting. A conference of international experts in SMA, including members of the original Hammersmith team, met to evaluate and suggest modifications to the scale that would limit institutional biases and rely as little as possible upon individual instruction or privately held

understandings that were not expressly part of the written instructions for the scale. We then sought to establish the test–retest reliability of the scale, and the extent to which scores are stable over the anticipated duration of a treatment trial, in order to estimate power of the test for purposes of later study design. We hoped to evaluate reliability of the measure in children younger than 30 months of age in light of recent evidence demonstrating progressive denervation in SMA 2 children prior to this age [27]. Finally, we wanted to evaluate the instrument in those with popliteal* angles of greater than 20°, which could theoretically impair scoring, function and thus reliability on some of the test items.

2. Materials and methods

The study consisted of 3 phases: phase 1— adaptation of the Hammersmith functional motor scale for children with SMA to ascertain objective and reliable use of the functional motor scale in a research setting and establishment of content validity of the adapted scale, the modified Hammersmith functional motor scale for children with SMA; Phase 2— evaluator training; and Phase 3— establishment of inter- and intrarater reliability, test–retest stability and discriminative validity of the modified scale.

2.1. Phase I: scale adaptation for use in a research setting, establishment of content validity

2.1.1. Scale adaptation for use in a research setting

In this phase, the original Hammersmith functional motor scale for children with SMA was adapted for research use. Concrete operational definitions were developed and scoring was clarified for the modified scale. Several multidisciplinary consensus meetings were held to develop operational definitions and clarify scoring criteria to minimize potential ambiguities in the administration and scoring of the test. Discussions focused on item interpretation and relevance, scoring criteria, and test procedures, including test environment. Participants in the consensus meetings included Project Cure SMA (PC-SMA) team members, Marion Main, physiotherapist, who first documented reliability and validity of the Hammersmith functional motor scale for children with SMA, and a pool of occupational therapists (OT's), physical therapists (PT's), and physicians who had experience with patients with SMA and outcome measure development.

*The popliteal angle has been defined differently by different investigators, though the principle of it representing a measure of hamstring restriction is the same. It has been defined by some as the angle subtended by the popliteal fossa; which is the maximum angle defined by the femur and tibia, when the hip is flexed to 90 degrees. Other authorities, favored in this report, define the popliteal angle as the complement of that angle, the angle defined by how much short of straight (180 degrees) the maximum knee extension is.

Once the pilot study was initiated an expert panel of consultants, through review of videotapes and discussions with clinical evaluators, periodically re-examined the rating scale. Refinements to test procedures, wording of operational definitions and item scoring were performed to ensure continued objectivity and test clarity. A procedure and direction manual to train evaluators was developed. See Figs. 1 and 2. A standardized video protocol to be used during each assessment was developed and instituted.

During initial scale modification, the item order of the original scale (Table 1) was left unchanged. The original Hammersmith functional motor scale for children with SMA test items ordered in a manner that was determined to be hierarchical in performance outcome [10]. This ordering required multiple position changes throughout the test. However, PC-SMA clinical evaluators consistently found that children became fatigued during testing and behavioral compliance often fluctuated secondary to frequent position changes during testing. Therefore, after a trial period of use (1 year) item order was changed to decrease fatigue and undue stress on the children during testing. Environmental stimuli, including toys, appeared to also influence behavioral performance. Re-ordering the items to limit position changes (Table 2) and institution of procedures to decrease environmental distractions improved behavioral performance and efficiency of test administration.

The changes made were not intended to alter the context of the test, but rather to assure for ease of use and standard administration and scoring procedures in a research setting. A preliminary adaptation of the scale for research use was completed in January 2002. After trial use of the modified Hammersmith functional motor scale for children with SMA for a period of 1 year, it was further modified to its current and final version in June 2003.

2.1.2. Establishment of content validity

Expert clinicians experienced in working with children with SMA reviewed the modified scale for content validity. All reviewers believed the items to be sensitive to functional change over time in children with SMA and to be useful to detect gross motor change in typically developing children. Items were also reflective of antigravity muscle strength. A review of the developmental literature suggests that all items on the scale are achieved by normal toddlers prior to the age of 20 months [28–30]. Main [10] suggests that children older than 30 months of age who achieve less than a full score on the original Hammersmith functional motor scale for children with SMA should be considered to have functional motor impairment.

2.2. Phase II: evaluator training and standardization of methods

Each evaluator underwent training in use of the modified scale at his/her clinic site or at a group site prior to initiating clinical assessment of patients. Training consisted of

Pulling or pushing with hands means they are using arms to initiate, complete or walk self over in roll vs. free rolling with other body muscles as they should. They are basically inching over by grabbing and pulling or pushing on surface with hands; they do this the weaker they are. Pulling on shirt sleeves/ toys may occur. Limit by no toys and short sleeves.

Item 8 and 10: Rolling prone to supine:

Score 2 if child rolls completely from prone to supine.

Score 1 if child rolls completely but pushes with hand to get over.

Score 0 if child is unable to achieve task.

***Stronger children may try to hold arms up or use arms in airplane fashion or in some fashion not stated above. If you know that the child is sufficiently strong to roll without arms please advise child to do so. If he/she still rolls with arms up in air- may score a 2 as long as they are not using arms to complete the roll.

Start position:



Pushing with hands, score 1 if completes roll.



Pushing with hands, score 1 if completes roll.



Fig. 1. Sample of test item criterion with operational definitions, scoring criteria and scoring examples from the modified-Hammersmith functional motor scale for children with SMA test manual for test item ‘rolling prone to supine’.

lecture, review, and videotape assessment to establish consistency in administration and scoring of the test. During the training each evaluator practiced using the scale by assessing children diagnosed with non-ambulatory SMA types 2 and 3. Standardization of methods was discussed and each evaluator was provided with an established set of criteria that would assure methods standardization from clinic to clinic. In addition, equipment needs and filming procedures were defined. There was adequate time for questions and answers to assure that each evaluator felt comfortable with the use of the scale after training.

Prior to the training session each evaluator scored 6–10 video CD’s of modified Hammersmith functional motor scale assessments. After training they rescored the same video CD’s to determine if training improved performance in use of the scale. After the training they also scored a set of training videos and these scores were compared with those of experts to assure all met a criterion of 90% agreement with expert scorers.

After each clinic started their data collection video CD’s were reviewed by outcome consultants to assure that

pre-established standardization criteria had been met. CD review and consultation with clinical evaluators continued until each clinical evaluator at each site satisfactorily achieved standardized performance criteria.

2.3. Phase III: reliability, test–retest stability over time, and discriminant validity

2.3.1. Instrument

The modified Hammersmith functional motor scale for children with SMA consists of 20 items (Table 2), each scored on a 3-point ordinal scale (2 for unaided, 1 for assistance, 0 for inability). The total test score can range from 0 if all the items are failed to 40 if all the items are achieved. All items are administered without thoracic or lower extremity orthoses. The test can be completed in 15–30 min.


2.3.2. Study participants

Eligible subjects all had a diagnosis of childhood spinal muscular atrophy (non-ambulatory SMA types 2 or 3) genetically confirmed by standard tests demonstrating


Item 1: Floor/chair sitting
 Child should be sitting on floor, chair, plinth or mat table. Document start position as chair/plinth or floor (frog) sitting. No back support allowed if sitting on floor or plinth. Ask child to raise hands and sit without hand support. Cannot place/prop hands on lap or other body part.

If raises both hands and maintains for 3 seconds- **score 2**.
 If requires one hand support to maintain- **score 1**.
 If requires 2 hand support- **score 0**.

Score 2



Score 1



Score 0




Fig. 2. Sample of test item criterion with operational definitions, scoring criteria and scoring examples from the modified-Hammersmith functional motor scale for children with SMA test manual for test item 'floor/chair sitting'.

homozygous deletion of SMN1 and clinically determined by the child's ability to maintain a sitting position when placed. All patients admitted to the study were between 9.53 months and 12 years of age. They were in good health with the exception of SMA and did not require BiPAP greater than 12 h per day. All enrolled subjects/guardians provided informed consent per institutional review board (IRB) standards at the institution where they participated. Children were excluded from the study if they had orthopedic

restrictions (e.g. internal spinal fusion) or equipment requirements that compromised use of the planned outcome assessment. They were also excluded if they had participated in a treatment trial for SMA in the 3 months prior to this trial, or planned on enrolling in any other treatment trial during the duration of this trial. All evaluators were licensed physical therapists with pediatric experience. Two consultant pediatric therapists (an OT and a PT) were responsible for training all evaluators. The consultant therapists

Table 1
 Item order original Hammersmith functional motor scale for children with SMA

1.	Frog/chair sitting no hand support
2.	Long sitting. No hands
3.	1/2; Roll from supine, both ways
4.	Touches one hand to head (R/L) (in sitting)
5.	Touches 2 hands to head (in sitting)
6.	Rolls prone to supine over R
7.	Rolls prone to supine over L
8.	Rolls supine to prone over R
9.	Rolls supine to prone over L
10.	Gets to lying from sitting (safely, not accidentally)
11.	Achieves prop on forearms-head up
12.	Lifts head from prone (arms down by sides)
13.	Achieves four point kneeling-head up
14.	Achieves prop on extended arms-head up
15.	Gets to sitting from lying through side lying
16.	Crawls
17.	Lifts head from supine
18.	Stands holding on with one hand
19.	Stands independently: count > 3
20.	Takes > 4 steps unaided

Table 2
 Item order modified-Hammersmith functional motor scale for children with SMA

1.	Frog (floor)/chair sitting no hand support
2.	Long sitting, no hands
3.	Raises one hand to ear level (R/L) (in sitting)
4.	Raises 2 hands to ear level (in sitting)
5.	Gets to lying from sitting (safely, not accidentally)
6.	Lifts head from surface in supine
7.	1/2; Roll from supine, both ways
8.	Rolls prone to supine over R
9.	Rolls prone to supine over L
10.	Rolls supine to prone over R
11.	Rolls supine to prone over L
12.	Lifts head from prone (arms down by sides)
13.	Achieves prop on forearms-head up
14.	Achieves prop on extended arms-head up
15.	Achieves four point kneeling
16.	Crawls on hands and knees
17.	Gets to sitting from lying through side lying
18.	Stands holding on with one hand
19.	Stands independently: count > 3
20.	Takes > 4 steps independently

developed and distributed a training manual and a CD for the modified Hammersmith functional motor scale for children with SMA. In addition, the consultant therapists were responsible for assuring standardization of testing methods by each evaluator and assuring that videotape procedures were consistently followed.

2.3.3. Intrarater reliability (in person)

A volunteer sample of 13 children with type 2 SMA (age range 2.2–9.7 years) was assessed by two clinical evaluators. One evaluator had previous experience in 20 Hammersmith assessments and the second evaluator had completed greater than 200 Hammersmith assessments. Each child was assessed twice by the same evaluator over a 2 day period. One evaluator examined seven subjects, the other examined six. Evaluators were blinded to each session's test results. The order of testing was randomized each day to minimize bias induced by fatigue of the rater and test order. Intrarater reliability of the modified Hammersmith functional motor scale for children with SMA was tested by comparing days 1 and 2 test results on the scale. Normality of the scores was checked and a linear regression analysis was performed to generate a Pearson correlation coefficient.

2.3.4. Interrater reliability/score–rescore reliability (of videos)

Over a two-year period a total of 44 children with type 2 SMA (age range 9.53 months–12 years) were assessed at two sites (Utah, Montreal) using the modified scale. All assessments were videotaped. Interrater reliability was assessed by review of 14 randomly selected videos by four blinded clinical evaluators from four different clinical sites. All evaluators had been trained in the use of the tool and had experience using the scale. Experience in use of the scale varied between evaluators from one assessment to greater than 200 total assessments. Interrater reliability was assessed using the intraclass correlation coefficient, assuming the raters were a representative sample of all raters and all raters evaluated the same subjects [31].

2.3.5. Intrarater scorer reliability (of videos)

To determine the scoring consistency of individual raters, four raters (with previous experience ranging from 1 to >200 assessments) scored the videotapes used for the interrater study, and then rescored those same videotapes 8–12 weeks later. To determine if the same evaluator was able to score consistently the same performance of the same child on the scale, the strength of agreement between repeat scorings of the same videotape by the same therapist was examined. Tapes were randomized to minimize bias and raters did not have access to previous scores. This data provides an evaluation of whether the reliability for scoring from video is similar to live evaluation. Since there are four raters the intraclass correlation coefficient was used assuming the raters were fixed (ICC (3,1)) [31] providing a measure intrarater reliability.

2.3.6. Test–retest reliability/test stability over time

Test–retest reliability measures the ability of a score on the scale to remain constant when there is no assumed change in the property that is being measured [32]. We hypothesized that there would be no change in modified Hammersmith functional motor scale scores during a baseline period of 3–6 months, as we expected little change in functional performance based on clinical experience, as well as data from an extensive natural history database. To determine if the scale was stable over a baseline period each subject was evaluated with the scale on admission to the study (T0), after 12 weeks (T1), and after 24 weeks (T2). The site's clinical evaluator performed each assessment without access to previous results to prevent any duplication or bias. We also hypothesized that reliability could be compromised by age less than 2 years, and by range of motion (ROM) deficits due to significant knee contractures (defined as popliteal angles greater than 20°).

Thirty-seven children with SMA type 2 participating in a larger natural history trial at two sites who had two baseline visits within a 6-month period were included. Three groups were differentiated based upon our pre-existing concerns: (Group A) 14 children who met initial eligibility criteria thought to be essential to modified Hammersmith functional motor scale for children with SMA reliability (>2 years of age and less than or equal to 12 years of age, popliteal angles <20°); (Group B) 7 children less than 2 years of age with popliteal angles <20°; (Group C) 14 children between 2 and 12 years of age with popliteal angles >20°. Performance on the scale was assessed at times T0, T1 and T2 by an evaluator blinded to previous results. Performance scores at T0, T1 and T2 were compared to determine test–retest stability. Each group's scores were compared to determine if pre-existing concerns (age <2 years of age or popliteal angles >20°) affected test–retest stability.

2.3.7. Discriminant validity

It was assumed that the modified Hammersmith functional motor scale for children with SMA should be able to detect a difference between types 2 and 3 SMA patients. Scores on the scale for type 2 patients were compared to scores for children with type 3 SMA to determine the discriminant validity of the instrument. Analysis of variance was used to evaluate differences between these groups.

3. Results

3.1. Measurements and main results

3.1.1. Intrarater reliability

There were 26 total assessments completed by two evaluators. The average Hammersmith score was 18.7 with a median of 17.5. The scores were normally distributed with a minimum of 3 and a maximum of 40. A limited number of

weak or minimally ambulatory SMA 3 subjects were included to help ensure reliability at the extreme ends of the functional spectrum for this cohort. Using a linear regression analysis we examined the relationship between the first and second test and the reliability coefficient was 0.99 indicating excellent reliability within rater.

3.1.2. Interrater reliability/score–rescore reliability

There were 14 video assessments scored by four raters. The scores ranged from 4 to 40 with a mean of 16.7 and a median of 16. The intraclass correlation coefficient (ICC (2,1)) demonstrated interrater reliability of 0.953 at a 95% confidence interval (0.913, 0.982).

3.1.3. Intrarater scorer reliability (of videos)

The four reviewers each reviewed every video twice with a separation of 8–12 weeks and in random order. The values ranged from 4 to 40, with a mean of 16.7 and median of 16. The intraclass correlation coefficient (ICC (3,1)) was 0.986 with a 95% confidence interval (0.962, 0.994) indicating that the scoring of the videos was very consistent within each reviewer. Reviewers 1 and 3 had the least experience and reviewers 2 and 4 had the most. The scores on the first evaluation of all patients were compared between the least and most experienced reviewers indicating no difference between reviewers by experience ($P=0.65$, ANOVA).

3.1.4. Test–retest reliability/test stability over time

Group mean scores at T0 and T1 are summarized in Fig. 3. Individual scores for all patients in each group are summarized in Fig. 4.

Group A: the interval between observations was 3.4 ± 1.0 SD (range 1.8–4.8) months. First (mean 18.7 ± 10.3 , range 6–40) and second (mean 19.1 ± 10.8 , range 6–40) scores were normally distributed. Linear regression analysis yielded an r -square value of 0.917 and a paired t -test indicated the change of 0.36 was not statistically significant at $P=0.67$. This

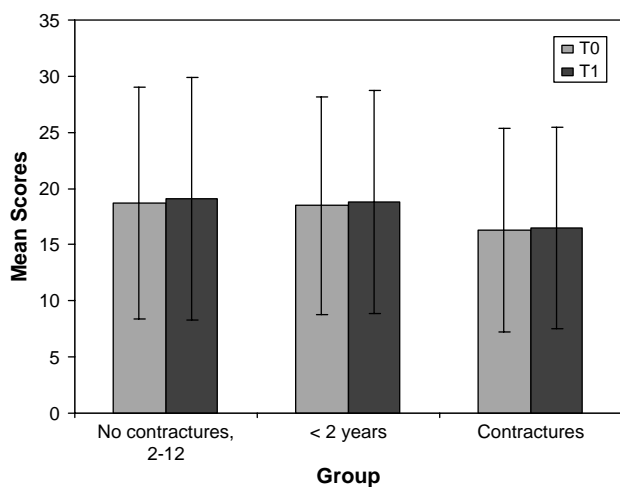


Fig. 3. Test–retest stability of the modified-Hammersmith functional motor scale for children with SMA. Mean scores for each group at T0 and T1.

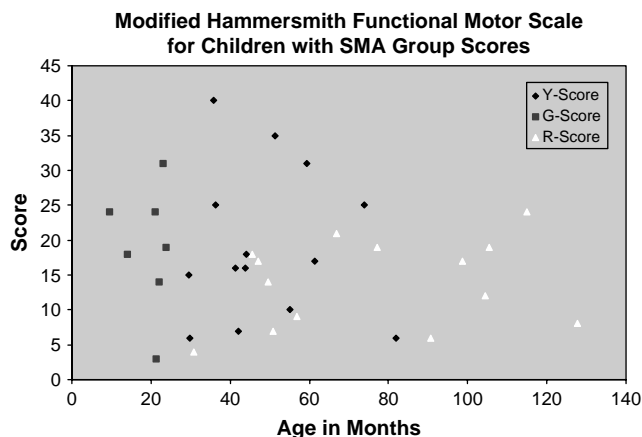


Fig. 4. Summary of modified-Hammersmith functional motor scale for children with SMA scores for patients in each identified subgroup. Y-Score- all patients between 2 and 12 years with popliteal angles $< 20^\circ$, G-Score- all patients under 2 years of age, R-score- all patients between 2 and 12 years with popliteal angles $> 20^\circ$.

indicates that the test is reliable, if given within 6 months for the type 2 patients that met all original inclusion criteria.

Adding the younger children of Group B to Group A showed no deterioration of reliability. Average time between observations was 3.0 months; range was 0.7–4.8 months with a SD of 1.3 months. The average first score used was 18.5, range 2–40, SD 9.7. The second score used had an average of 18.8, range 3–40, SD 9.9. All scores were normally distributed. Linear regression analysis yielded an r -square value of 0.91 and a paired t -test indicated the change of 0.24 was not statistically significant at $P=0.78$. This indicates that the test is reliable, if given within 6 months for the type 2 patients after lowering the entry age limit. The mean age of the original group was 48.98 months with a range of 29.46–81.9 months. The younger group had a mean of 20.02 months with a range of 9.53–24.59 months.

Adding the children of Group C with greater popliteal contractures similarly demonstrated no deterioration of reliability (interval 3.4 ± 1.06 , range 1.8–5.98 months; first score mean 16.3 ± 9.1 , range 2–40, second score mean 16.5 ± 9.0 , range 4–40; r^2 0.902, paired t -test of 0.25 yielded a non-significant $P=0.65$). An ANOVA was used to test for differences in the ages between children without contractures and children with contractures. The average age for the contracture patients was 75.3 months and 51.5 months for the eligible patients without contractures ($P=0.02$). There was no difference in score between the children with and without contractures ($P=0.56$) as determined by an ANOVA. For this an ANOVA assessed the first observation in from the pair of observations used in the test–retest reliability analysis.

Finally, all three groups were compared to determine if there was any difference in score. An ANOVA was used and exhibited no difference between the three groups, $P=0.32$.

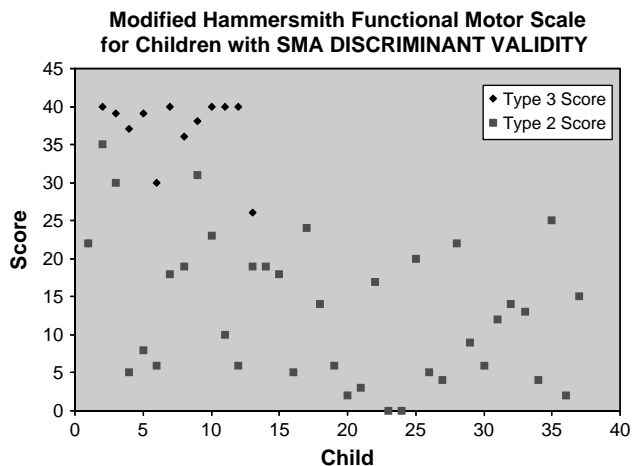


Fig. 5. Distribution of scores for types 2 and 3 patients used in discriminative validity analysis. Of note: both type 3 children with lower scores demonstrated significant issues with behavior and cooperation at time of testing.

3.1.5. Discriminant validity

There were 37 patients with non-ambulatory SMA types 2 and 3 with modified Hammersmith functional motor scale scores. The average score was 13.3 (standard error of measure (SEM), 1.53). There were 13 patients with type 3 SMA with modified Hammersmith functional motor scale scores. The average scale score was 35.9 (SEM 1.67). Patient scores are summarized in Fig. 5. These two findings are statistically, significantly different at $P < 0.0001$ using the Wilcoxon two-sample test which was used because the scores were not normally distributed.

4. Discussion

The modified Hammersmith functional motor scale for children with SMA appears to be well suited for use as a primary outcome measure in treatment trials of young non-ambulatory children with SMA who are able to sit unsupported. Although, it theoretically could be used in weaker type 2 subjects who have lost the ability to sit, and in stronger SMA subjects who have the ability to stand or even take a limited number of steps without support, these populations are less ideal for use of the scale as it currently stands. For a cohort of non-ambulatory SMA sitters, the intrarater and interrater reliability are high. Importantly, we have demonstrated that modified Hammersmith functional motor scale for children with SMA scores are stable over a 6 month period. In addition, scores assessed in children as young as 24 months, and in those with hamstring contractures manifesting a popliteal angle greater than 20° appear to be as reliable as in older children and children without contractures.

In order to have a relatively homogenous cohort of patients who could all be tested on the same scale, we initially proposed inclusion of non-ambulatory children with SMA

types 2 and 3 between 2 and 12 years of age. Although it has been documented that non-disabled children over 29 months of age can consistently achieve a full score (39–40) on the original Hammersmith scale we did not exclude children younger than 30 months, as we wanted to test the hypothesis that the tool could be reliable in those younger than 30 months as we assume that younger children could potentially achieve a greater benefit if treated as early as possible in the course of their disease. Natural history suggests that a plateau in motor development probably occurs early in the course of the disease. We wanted to examine the motor stability in younger children to determine the value of this motor scale in assessing functional outcomes in this younger cohort who could potentially achieve greatest benefit in clinical trials. Children older than 12 years were excluded, as after this age several complications, such as severe scoliosis and contractures are more frequent. Reliability is maintained when we expand eligibility criteria to younger children and to those with popliteal angles greater than 20° . These results will be important to the design of planned treatment trials of SMA.

Intrarater reliability of the modified Hammersmith functional motor scale for children with SMA in live patients, and interrater reliability of videotaped sessions, is very high. These findings suggest that the scale is a useful instrument for use in multi-site collaborative treatment trials of non-ambulant children with types 2 and 3 SMA between 2 and 12 years of age. In addition, we have demonstrated reliability in some children as young as 9 months of age. Inter rater reliability looked at reliability of scoring alone as we used videotapes. Further work needs to be done to look at reliability of testers administering and scoring the test rather than just scoring alone.

The results of this study support the use of the modified Hammersmith functional motor scale for children with SMA as a valid and reliable measure of change in gross motor function in non-ambulatory children with SMA as young as 2 years of age. The scale demonstrated non-significant variation under stable conditions. In addition, in an open label valproic acid treatment trial in this same patient population, it appears to detect significant change when change was believed to have taken place. It also demonstrated the potential to pick up varying levels of change. This indicates that the scale has the potential to be responsive to change in clinical trials of children with SMA types 2 and 3, who are non-ambulant.

Overall the high reliability estimates are reflective of the heterogeneous population of the children with SMA 2 assessed in this study. Reliability estimates might be lower if calculated in a more homogeneous group. On the other hand, the group of children used in this study was representative of the population of non-ambulatory SMA sitters, primarily SMA 2 subjects, for whom this test was developed.

Putative outcome measures for a clinical trial of SMA are complicated by the difficulty in establishing a highly reliable measure of muscle power in young children.

Functional scales of power have been limited by the perceived need to design an instrument that encompasses sensitivity to the whole range of disease severity. An inevitable consequence of this effort has been that ceiling and floor effects limit the ability to detect change, while reliability is compromised by difficulties with inter-observer variability [3]. We did not find significant floor and ceiling effects in the chosen cohort and had good inter-observer reliability. However, future trials may be best served by an appropriately extended scale to minimize potential floor and ceiling effects and capture performance changes in the weaker and stronger ends of the spectrum of children with SMA.

A number of sources of variability may reduce the reliability of a test. These include: variation due to raters, subjects, the environment and the test itself. Several steps were taken to maximize true responses and minimize variability in our study. To eliminate interrater variability in the study the same PT administered the scale on both occasions. To reduce intrarater variation all the evaluators were trained to use the scale. To put the child at ease the therapist kept the environment as consistent as possible including room and time of day. In addition, as many children traveled long distances to a clinic site, travel and fatigue may influence behavioral performance. We have tried to control for this by scheduling and testing during a time of day that each child is well rested and nutritionally satisfied.

Some children's performances were atypical. Children younger than 30 months appeared to have some difficulty attending and cooperating during testing. This may be age typical behavior and has been noted previously by the test developers. Reliability and sensitivity data for all children younger than 30 months was assessed and compared to reliability and sensitivity data for all children above 30 months of age. It was determined that age was not a factor. However, in the youngest children, delayed achievement of motor milestones could theoretically lead to improvements in scores that are not clearly treatment related, but a function of developmental maturation. Although, we did look at a group of children less than 24 months of age further assessment of this age group should be undertaken as our study population for this group was small. Developmental maturation may be proposed as a confounding variable in testing those younger than 24 months; however our cohort remained stable over the period of the study which suggests, as does natural history, that development plateaus early in the course of the disease. Additional studies of children less than 24 months will need to be performed to best assess the most appropriate use of this instrument in that population.

Additional difficulties encountered by evaluators during initial use of the modified Hammersmith functional motor scale for children with SMA in a research setting included the need to periodically clarify test criteria and operational definitions. The original Hammersmith functional motor scale for children with SMA was modified in order to

improve its research applicability and to assure objective use and reliability in a multi-site clinical trial. By having objective definitions of items and a standardized scoring system with a test manual observer variation was minimized.

Literature suggests that rater training for observational clinical instruments is important for both administration and scoring [33,34] and rater training [35] and familiarity with a test instrument have been found to affect the reliability of clinical measures and the consistency of scoring [36,37]. Training is often associated with high interrater reliability in observational clinical instruments. Use of a test manual with clear operational definitions and photographs as well as onsite training was undertaken to minimize the use of personal interpretation and clinical experience while administering and scoring the modified scale. Training provided raters with improved knowledge of the clinical measure and its' operational definitions which facilitated optimized administration in a standardized manner. It also provided raters with improved scoring experience through use of immediate feedback, modeling, information and practice, which may have enhanced their reliability. However, onsite training incurs increased use of resources and time. It would thus be of value to further explore the effect of rater training and rater familiarity with the modified scale on interrater reliability, internal consistency and standard error of measure to determine its overall effect and necessity.

In summary, this study supports the use of the modified Hammersmith functional motor scale for children with SMA to assess change in the non-ambulatory child with SMA, and adds objectivity to ascertain standardized and reliable use in multi-site clinical trials. Continued exploration of items that may detect change in weaker and stronger children who function outside the boundaries of the current scale is ongoing. Items that further assess fine motor and ADL abilities, as well as timed motor tests that will assess endurance and power could be considered as add-on modules to the current scale. A more comprehensive scale may be more sensitive to the heterogeneous spectrum of motor abilities noted in this population and may allow us to assess change in an even more sensitive and efficacious manner. However, we must remain sensitive to the variability induced by excessive fatigue in this population. Standards for use of functional outcomes must demonstrate a high degree of objectivity and reliability in order to minimize potential bias and optimize standardization of use across sites in multi-site clinical trials.

Acknowledgements

We would like to thank the children and their families who gave generously of their time and effort. We would also like to acknowledge the clinical evaluators for their time and expertise, particularly Janine Wood, Karine Nolet, Natalie

Mellem, Liz Bollman, and Jill Hartman. We would also like to thank Mark Wride for all of his assistance and his expertise in development and coordination of the database and Marion Main for sharing her clinical expertise in use of the original scale. We are also grateful for the inspiring collaboration of all other Project Cure Team members who participated in this project. Families of Spinal Muscular Atrophy provided funding for this study as part of the Project Cure SMA initiative. Additional funding was provided to KJS by the Muscular Dystrophy Association, and the American Academy of Neurology and Spinal Muscular Atrophy Foundations.

References

- [1] Andreassi C, Angelozzi C, Tiziano FD, et al. Phenylbutyrate increases SMN expression in vitro: relevance for treatment of spinal muscular atrophy. *Eur J Hum Genet* 2004;12(1):59–65.
- [2] Brahe C, Vitali T, Tiziano FD, et al. Phenylbutyrate increases SMN gene expression in spinal muscular atrophy patients. *Eur J Hum Genet* 2005;13(2):256–9.
- [3] Crawford TO. Concerns about the design of clinical trials for spinal muscular atrophy. *Neuromuscul Disord* 2004;14(8–9):456–60.
- [4] Beenakker EA, van der Hoeven JH, Fock JM, Maurits NM. Reference values of maximum isometric muscle force obtained in 270 children aged 4–16 years by hand-held dynamometry. *Neuromuscul Disord* 2001;11(5):441–6.
- [5] Merlini L, Solari A, Vita G, et al. Role of gabapentin in spinal muscular atrophy: results of a multicenter, randomized Italian study. *J Child Neurol* 2003;18(8):537–41.
- [6] Miller RG, Moore DH, Dronsky V, et al. A placebo-controlled trial of gabapentin in spinal muscular atrophy. *J Neurol Sci* 2001;191(1–2):127–31.
- [7] Merlini L, Mazzone ES, Solari A, Morandi L. Reliability of hand-held dynamometry in spinal muscular atrophy. *Muscle Nerve* 2002;26(1):64–70.
- [8] Merlini L, Bertini E, Minetti C, et al. Motor function-muscle strength relationship in spinal muscular atrophy. *Muscle Nerve* 2004;29(4):548–52.
- [9] Sloan C. Review of the reliability and validity of myometry with children. *Phys Occup Ther Pediatr* 2002;22(2):79–93.
- [10] Main M, Kairon H, Mercuri E, Muntoni F. The Hammersmith functional motor scale for children with spinal muscular atrophy: a scale to test ability and monitor progress in children with limited ambulation. *Eur J Paediatr Neurol* 2003;7(4):155–9.
- [11] Chung BH, Wong VC, Ip P. Spinal muscular atrophy: survival pattern and functional status. *Pediatrics* 2004;114(5):e548–e53.
- [12] Iannaccone ST. Outcome measures for pediatric spinal muscular atrophy. *Arch Neurol* 2002;59(9):1445–50.
- [13] Rudnik-Schoneborn S, Hausmanowa-Petrusewicz I, Borkowska J, Zerres K. The predictive value of achieved motor milestones assessed in 441 patients with infantile spinal muscular atrophy types II and III. *Eur Neurol* 2001;45(3):174–81.
- [14] Russman BS, Buncher CR, White M, Samaha FJ, Iannaccone ST. Function changes in spinal muscular atrophy II and III. The DCN/SMA group. *Neurology* 1996;47(4):973–6.
- [15] Moxley 3rd RT. Functional testing. *Muscle Nerve* 1990;13:S26–S9.
- [16] Vignos Jr PJ, Spencer Jr GE, Archibald KC. Management of progressive muscular dystrophy in childhood. *J Am Med Assoc* 1963;184:89–96.
- [17] 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(3):186–97.
- [18] Brooke MH, Griggs RC, Mendell JR, Fenichel GM, Shumate JB. The natural history of Duchenne muscular dystrophy: a caveat for therapeutic trials. *Trans Am Neurol Assoc* 1981;106:195–9.
- [19] Scott OM, Hyde SA, Goddard C, Dubowitz V. Quantitation of muscle function in children: a prospective study in Duchenne muscular dystrophy. *Muscle Nerve* 1982;5(4):291–301.
- [20] 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(3):119–34.
- [21] Cedarbaum JM, Stambler N, Malta E, et al. The ALSFRS-R: a revised ALS functional rating scale that incorporates assessments of respiratory function. BDNF ALS study group (phase III). *J Neurol Sci* 1999;169(1–2):13–21.
- [22] Cedarbaum JM, Stambler N. Performance of the amyotrophic lateral sclerosis functional rating scale (ALSFRS) in multicenter clinical trials. *J Neurol Sci* 1997;152(Suppl 1):S1–S9.
- [23] Brooke MH, Fenichel GM, Griggs RC, et al. Clinical investigation of Duchenne muscular dystrophy. Interesting results in a trial of prednisone. *Arch Neurol* 1987;44(8):812–7.
- [24] Conolly AM, Schierbecker J, Renna R, Florence J. High dose weekly oral prednisone improves strength in boys with Duchenne muscular dystrophy. *Neuromuscul Disord* 2002;12(10):917–25.
- [25] Fenichel GM, Florence JM, Pestronk A, et al. Long-term benefit from prednisone therapy in Duchenne muscular dystrophy. *Neurology* 1991;41(12):1874–7.
- [26] Fenichel GM, Griggs RC, Kissel J, et al. A randomized efficacy and safety trial of oxandrolone in the treatment of Duchenne dystrophy. *Neurology* 2001;56(8):1075–9.
- [27] Swoboda KJ, Prior TW, Scott CB, et al. Natural history of denervation in SMA: relation to age, SMN2 copy number, and function. *Ann Neurol* 2005;57(5):704–12.
- [28] Piper MC, Pinnell LE, Darrah J, Maguire T, Byrne PJ. Construction and validation of the Alberta infant motor scale (AIMS). *Can J Public Health* 1992;83(Suppl 2):S46–S50.
- [29] Egan DF, Illingworth RS, Mac Keith RC. Developmental screening 0–5 years, clinics in developmental medicine. London: Spastic International Medical Publications; 1969.
- [30] Folio MR, Fewell RR. Peabody developmental motor scales: examiner's manual. 2nd ed. Austin, TX: Pro-Ed; 2000.
- [31] Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. *Psychol Bull* 1979;86:420–8.
- [32] Portnoy LG, Watkins MP. Foundations of clinical research: applications to practice. 1st ed. Norwalk, CT: Appleton & Lange; 1993.
- [33] Castorr AH, Thompson KO, Ryan JW, Phillips CY, Prescott PA, Soeken KL. The process of rater training for observational instruments: implications for interrater reliability. *Res Nurs Health* 1990;13(5):311–8.
- [34] Haller KB. Interrater reliability: essential for research and practice. *MCN Am J Mater Child Nurs* 1987;12(1):78.
- [35] Ryan JW, Phillips CY, Prescott PA. Interrater reliability: the underdeveloped role of rater training. *Appl Nurs Res* 1988;1(3):148–50.
- [36] Backman C, Mackie H. Arthritis hand function test: inter-rater reliability among self-trained raters. *Arthritis Care Res* 1995;8(1):10–15.
- [37] Washington CC, Moss M. Pragmatic aspects of establishing interrater reliability in research. *Nurs Res* 1988;37(3):190–1.