J Inherit Metab Dis (2007) 30:768–782
DOI 10.1007/s10545-007-0595-x

ORIGINAL ARTICLE

A severity scoring tool to assess the neurological features of neuronopathic Gaucher disease

E. H. Davies · R. Surtees · C. DeVile · I. Schoon · A. Vellodi

Received: 26 February 2007 / Submitted in revised form: 29 May 2007 / Accepted: 7 June 2007 / Published online: 16 September 2007

Summary Type III Gaucher disease is one of the three recognized subtypes of Gaucher disease, an inherited deficiency of lysosomal glucocerebrosidase. Phenotypically there is a wide spectrum of visceral and neurological manifestations. Enzyme replacement is effective in managing the visceral disease; however, the neurological manifestations remain a more challenging obstacle. There is an unfulfilled need to reliably monitor neurological disease and its response to treatment. A severity scoring tool was developed through neurological domain identification, item generation and tool formation. Domain identification was established based on a retrospective single centre study (n=15) and a systematic review of publications. Forty-seven patients with neuronopathic Gaucher disease were then assessed using the tool to establish the clinical and statistical reliability of each domain. Judgement quantification of the tool was established through a process of content validity involving five European experts. Content validity is considered to be most effective when undertaken systematically. Concurrent validity and feasibility of the tool was also highlighted. This process allowed a revised and validated version of the tool to be developed.

Communicating editor: Georg Hoffmann

Competing interests: None declared

References to electronic databases: PubMed

E. H. Davies · A. Vellodi
Metabolic Medicine, Great Ormond Street Hospital NHS Trust
Biochemistry Research Group, Clinical & Molecular Genetics Unit, University College London Institute of Child Health, London, UK

R. Surtees · C. DeVile
Department of Neuroscience, Great Ormond Street Hospital NHS Trust, London, UK

I. Schoon
Social Science, City University, London, UK

E. H. Davies (✉)
Institute of Child Health, 30 Guildford Street,
London WC1N 1EH, UK
e-mail: E.Davies@ich.ucl.ac.uk

Abbreviations

BMT	bone marrow transplantation
CVI	index of content validity
ERT	enzyme replacement therapy
EWGGD	European Working Group of Gaucher Disease
HGP	horizontal gaze palsy
NGD	neuronopathic Gaucher disease
SARA	scale for the assessment and rating of ataxia
SIF	saccade initiation failure
SST	severity scoring tool

Introduction

Neurological involvement in Gaucher disease was first described by Rusca in 1921. 'Neuronopathic' forms are the rarest variant of Gaucher disease, with an estimated incidence of <1:100 000 live births. The management of neuronopathic Gaucher disease (NGD) is fraught with difficulty (Vellodi et al 2001). There is poor genotype–phenotype correlation. Bone marrow transplantation (BMT) and enzyme replacement therapy (ERT) have

altered the natural history of the disease, and hence the data that needs to be captured. Therefore, the requirements of a severity scoring tool (SST) are different now from what they would have been a decade ago.

Neuronopathic Gaucher disease is defined as the presence of neurological involvement in patients with biochemically proven Gaucher disease for which there was no explanation other than Gaucher disease. While dividing into subtypes is common practice, there is a general consensus that a continuous spectrum of phenotypes exists. Chronic NGD in this instance refers to patients who do not have the acute form. Detailed neurological manifestations of the disease were previously reported twenty-six years ago (Erikson et al 1980) before the introduction of BMT and ERT into clinical care (Fig. 1). At that time visceral disease was the major cause of death, and therefore neurological examination was difficult and possibly unreliable.

Monitoring of the disease severity and response to ERT in Gaucher type I (non-neuronopathic) is based on visceral disease and is quite reliable. However, neurological manifestations are difficult to quantify. Current monitoring varies, but may include neuropsychological assessments, brainstem auditory evoked potentials, and saccadic eye assessment, in addition to standard clinical examination. However, there is no standardized format for reporting the results of assessments. Disease-specific scales are very useful when the attributes of particular diseases or conditions require assessment, as they will usually be more sensitive (Bowling 2001).

Achieving validity and reliability of a measuring tool requires time and effort, which is a powerful reason for using existing scales when available (Bowling 2001). When developing a measurement for disease, the reliability, validity and sensitivity of the scale, the appropriateness of the instrument for the study population, and the acceptability of the instrument to the group under study need to be evaluated and considered.

Zimran (1989) developed a severity scoring index for Gaucher type I based on an Ashkenazi cohort of 53 patients. This is sporadically used to monitor patients, but it has limited value for capturing neurological involvement. All neurological involvement is calculated in one domain, and accounts for only 20 points out of a maximum possible score of 46. This does not allow for any change within the neurology to be captured. The Mainz Severity Score Index for Fabry Disease is another example where the response of a lysosomal storage disease to ERT has been standardized (Whybra et al 2004). The neurology of a lysosomal storage disorder has been captured in a tool by Steinfeld and colleagues (2002) for late-infantile neuronal ceroid lipofuscinosis. This clinical scoring system is disease-specific and was developed as a method to quantitatively describe the clinical course of the disease over many years. The paper demonstrates its value as a simple assessment system. However, oversimplification could lead to reduced sensitivity of the tool. A survey of the literature did not highlight a consistent way to monitor neurological manifestations of NGD. This remains an unmet need to date.

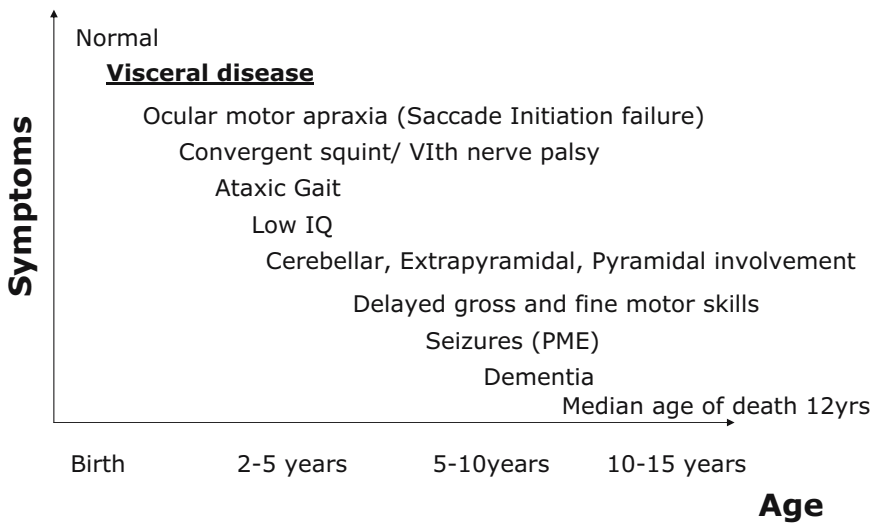

Fig. 1 Natural history of Gaucher disease type III according to Erikson et al (1980)

Research design

The primary objectives of study were:

- To develop a severity scoring tool (SST) that would capture all neurological features of NGD, numerically calculate the disease severity, and monitor the clinical course of the disease either naturally or in response to treatment.
- To evaluate the validity and reliability of the devised severity scoring tool.

Methodology

Initial data extraction

A retrospective study of 15 patients with a biochemically proven diagnosis of NGD at Great Ormond Street Hospital was performed. Data on all neurological features identified in the notes were initially extracted by one reviewer (first author) and then checked by a colleague. All neurological features were recorded as 'mild', 'moderate', or 'severe'. Disagreement was resolved by discussion between both reviewers.

Search strategies

A literature search was conducted to identify neurological manifestations of the disease. The value of utilizing published literature to inform about the natural history and presentation of lysosomal storage disorder has been demonstrated elsewhere (van den Hout et al 2003). Publications were identified via PubMed by a search for the terms 'neuronopathic Gaucher disease' AND/OR 'type III Gaucher disease'.

Inclusion and exclusion criteria for primary studies

Search was limited to publications in English after 1991, when ERT became available, in order to focus on the neurological features identified in the post-ERT era. Primary references were categorized by case studies and group cohorts. Selection criteria focused only on those that included clinical findings about patient status. Papers which only discussed mutation analysis or subclinical assessments (such as auditory brainstem testing, saccadic eye movement measurements, somatosensory evoked potential) were excluded as it was felt that they would not contribute to the evaluation and scoring of clinical presentations. Publications discussing neurological findings in type I and type II Gaucher disease were excluded. While it is acknowledged that case studies offer a wealth of valuable information it was felt that publication bias may favour the more severe spectrum of disease. A decision was therefore made to exclude individual case studies from analysis. In order to maximize benefit, only papers with cohorts of eight patients or more were included.

Tool formulation

The neurological features identified were incorporated as the domains to be included in the tool developed for pilot use in patients with NGD.

Establishing internal reliability

Item discrimination for a person-based test, as described by Rust and Golombok (1995), represents the extent to which the item is measuring the same concept as all of the other items in the questionnaire under the assumption that the construct is unidimensional. Items should only be selected for the final version of the scale on this basis. Cronbach's alpha is the standard measure of internal consistency (reliability), mathematically equivalent to the average of all possible split-half estimates. Cronbach's alpha is a measure of reliability that is a lower bound for the true reliability of an assessment tool. The computation of Cronbach's alpha is based on the number of items and the ratio of the average inter-item covariance to the average item variance.

Establishing validity

Validity is the extent to which that instrument measures what it is intended to measure. An SST of high validity will measure, as intended, the severity of neurological features in patients with NGD. There are numerous aspects of validity to be considered during tool development. Although not all of them are yet accounted for in this study, concurrent validity and feasibility is demonstrated.

Content validity

The legitimacy of content validity as a real type of validity has been repeatedly questioned (Lynn 1986). These challenges to the value and merit of content validity have arisen from the confusion of content validity and face validity, the unstandardized approaches to the determination of content validity,

and the previously unquantified nature of content validity.

Content validity is the determination of the content relevance of the elements of an instrument by the application of a two-stage process. Use of a two-stage process to determine and quantify content validity is fundamental to the validation of virtually all instrumentation. The assessment of content validity begins in the earliest development of an instrument. This two-stage process is referred to as development and judgement.

Judgement evaluation entails asking a specific number of experts to evaluate the validity of items individually and as a set. Content validity is based on consensual judgement by experts working independently. An index of content validity showing the proportion of agreement among judges is calculated. An index of content validity (CVI) using a 4-point ordinal scale was used to quantify content validity. A 4-point ordinal rating scale is preferable because it does not include the ambivalent middle rating common in odd-number rating scales, and provides sufficient delineated information upon which to calculate a meaningful CVI. The actual CVI is the proportion of items that received a rating of 3 or 4 by the experts.

The exact number of judges that might be used has not been established, but is likely to be a minimum of 3 and a maximum of 10 (Lynn 1986). The proportion of experts whose endorsement is required to establish content validity beyond the 0.05 level of significance as dictated by the number of experts involved was identified based on the formula devised by Lynn (1986). A CVI score of 0.80 or greater is considered to have excellent content validity. The rigour of the validation process can be greatly influenced by how experts are chosen and utilized for instrument development (Gibson et al 2006). Increasingly, content validity is being undertaken more systematically in the initial phase of instrument development (Gibson et al 2006; Richmond and Wright 2005; Woolery et al 2006). The importance of content validity during the entire process of instrument development cannot be underestimated, as an instrument can be reliable without being valid, but unreliable instrument cannot be valid (DeVellis 1991).

Concurrent validity

Concurrent validity can be tested by identifying the correlation between a new instrument and a previously validated instrument for measuring the same concept (Seong 2002). This is difficult in NGD as no previous assessment tool or 'gold standard' exists. In its absence this study aimed to determine whether the SST score predicts with a known clinical outcome. The scores of subgroups based on genotype and spleen status within the pilot cohort are therefore examined.

Feasibility

Although the reliability and validity of an assessment tool must be shown before an appraisal of its feasibility can be done with confidence, the issue of feasibility must be addressed before attempts at widespread adoption can take place. Local differences can give different perspectives of what constitutes feasibility. An assessment that is time and labour intensive diminishes the feasibility of widespread application.

Results

Findings in GOSH cohort

All patients were examined neurologically with detailed documentation between July 2003 and February 2004. There were 3 boys and 12 girls. The mean age was 9 years (range 2–20 years). Twelve (80%) were L444P homozygotes with one L279P/G243V, one L444P/E233D and one L444P/D409H. One patient had had a total splenectomy in preparation for BMT. Fourteen were on ERT. Two had had a partial splenectomy. Thirteen neurological features were identified in this cohort. Eleven features were relatively easy to quantify, and are described in Table 1 along with their severity and frequency. Fine motor skills and gross motor skills deficits were the additional two features present in all the patients at varying degrees of severity, but were difficult to quantify. All 15 had horizontal gaze palsy (HGP) or saccade initiation failure (SIF), which forms the basis for clinical distinction between type I and type III. Ataxia and pyramidal involvement appeared with greatest frequency, but there was considerable heterogeneity in presentation. Seven patients had kyphosis. Whether this is an aspect of bone disease or neurology is unclear, but it was retained as a neurological clinical feature in this study. Dementia and progressive myoclonic epilepsy as reported by Erikson et al (1980) were not identified in this cohort; but this could be due to the relatively young age of our cohort. Domain identifications and variables to populate domains are traditionally done through Delphi studies. In this instance, however, identified domains were compared with those extracted from the literature searches.

Table 1 Neurological features in the study cohort

Patient	HGP[a]	Seizures	Ataxia	Cerebellar	Pyramidal	Extra-pyramidal	Cognitive impairment	Ophthalmology (nerve palsy)	Swallowing difficulties	Speech	Spinal alignmnt	Age at diagnosis
1	✓✓		✓	✓	✓							1 year 4 months
2	✓✓✓		✓	✓	✓			✓✓	✓		✓✓	1 year 1 months
3	✓✓✓		✓✓		✓✓	✓✓		✓✓	✓	✓✓	✓✓	1 year
4	✓✓		✓		✓					✓		1 year 2 month
5	✓								✓		✓	1 year 4 months
6	✓✓		✓	✓	✓				✓		✓✓	1 year 5 months
7	✓		✓		✓		✓✓				✓	3 years
8	✓✓		✓	✓✓							✓	6 months (sibling)
9	✓✓		✓		✓							1 year 2 months
10	✓✓		✓✓	✓✓	✓✓	✓✓	✓✓		✓✓	✓	✓	3 months
11	✓✓	✓✓	✓	✓✓	✓✓							1 year 3 months
12	✓		✓		✓✓				✓✓✓			8 months
13	✓		✓		✓			✓		✓✓✓	✓	Data not available
14	✓✓			✓	✓							Data not available
15	✓✓			✓	✓					✓		Data not available
Total	15	1	12	8	13	2	2	3	6	5	8	

[a] HGP, horizontal gaze palsy.
✓=mild ✓✓=moderate ✓✓✓=severe.

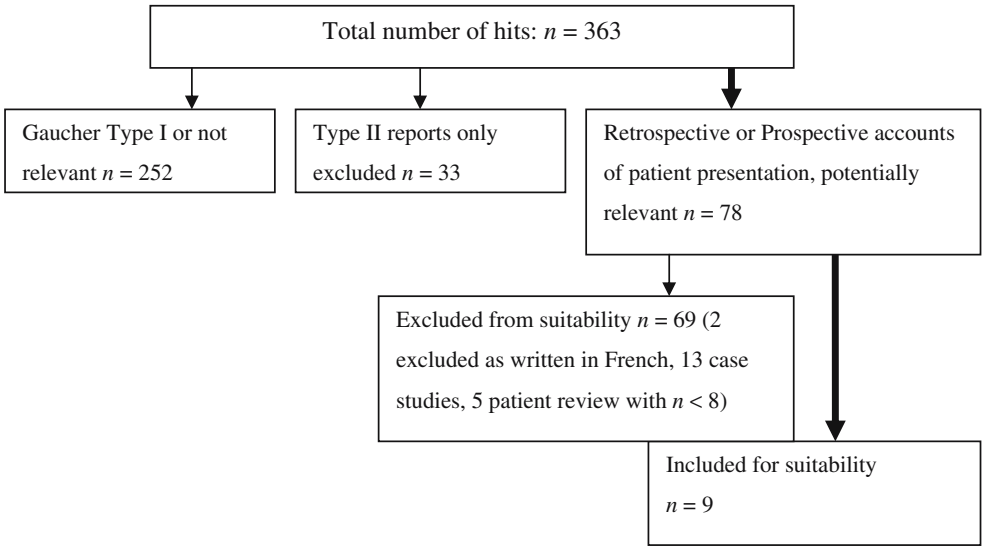

Fig. 2 Flowchart of study identification

Quantity and quality of studies identified

With the narrow search for neuronopathic Gaucher disease AND type III, the literature search yielded 363 references that were potentially relevant (Fig. 2). Nine are reported.

Presenting cohorts in selected publications

A total of 122 patients were reported. One paper was a 10-year follow-up of the same cohort (n=8) (Erikson et al 1995, 2006) and that of the original Norrbottnian cohort (Erikson et al 1980). A further two papers appeared to report different clinical data on the same Arab cohort (n=12) (Abrahamov et al 1995; Pasmanik-Chor et al 1996).

Fifty-eight per cent (n=59) of the 102 patients reported were L444P/L444P. Two cohorts had a 100% L444P homozygote cohort, one being Norrbottnian the other Japanese (Erikson et al 1995, 2006; Ida et al 1999). Three other cohorts also had a high L444P homozygote percentage at 75%, 72% and 62% (Altarescu et al 2001; Goldblatt et al 2005; Tylki-Szymanska et al 2006). These were primarily Australasian, Polish and caucasian. The Arab cohort reported by two authors was genotypically very different from the others; all were D409H homozygotes. A third paper (Park et al 2003) had only one L444P/L444P (6%). The neurological features identified in this paper, however, are very consistent with the other cohorts.

Fifty-seven per cent were male (n=54) with data on the sex of one cohort missing. Age at assessment for the cohorts was difficult to extrapolate; it appeared to range from 8 months to 40 years with a mean age of 14.3 years.

Six of the cohorts reported on the prescribed doses of ERT. This varied greatly from patient to patient and across cohorts, from 5 IU/kg per 2weeks (Ida et al 1999) to 120 IU/kg per 2 weeks (Altarescu et al 2001). Seven out of the 16 reported by Park and colleagues (2003) were not receiving any ERT. The spleen status was only sporadically reported.

Neurological features identified in the selected publications

The frequency of neurological findings was difficult to extrapolate from the publications in terms of individual patient frequency. Therefore, references to neurological findings are mainly based on complete cohorts.

Horizontal gaze palsy (HGP) or oculomotor apraxia was reported in 7 out of the 8 cohorts presented (Abrahamov et al 1995; Altarescu et al 2001; Erikson et al 1995, 2006; Goldblatt et al 2005; Ida et al 1999; Pasmanik-Chor et al 1996; Tylki-Szymanska et al 2006). This feature was documented as horizontal saccade involvement in the eighth (Park et al 2003). The primary focus of two papers was on visceral and cardiac disease, not neurology (Abrahamov et al 1995; Goldblatt et al 2005; Pasmanik-Chor et al 1996). Oculomotor apraxia was reported as the sole neurological sign. Interestingly two papers reported a total of five patients without the traditionally characteristically HGP associated with NGD (Erikson et al 1995, 2006; Ida et al 1999).

Epilepsy was reported in five cohorts, and classified with various specificity: myoclonus epilepsy, progressive myoclonus epilepsy, incipient cortical myoclonus, partial complex seizures, convulsions, seizures and

Table 2 Overview of neurological features as identified in the literature (n=110)

Author	n	Genotype	Mean age[range] (years)	Neurological features
Abrahamov et al (1995)[a]	12	100% D409H mutation	10.8[2–20]	Oculomotor apraxia
Altarescu et al (2001)	21	62% L444P/L444P	[8 months–35 years]	Autistic behaviour Epilepsy Horizontal supranuclear gaze palsy Low IQ
Erikson et al (1995)[b]	8	100% L444P/L444P	N/C	Ataxia Epilepsy Horizontal gaze palsy (ocular motor apraxia) Intention tremor Low IQ Progressive dementia Spastiscity Kyphosis
Erikson et al (2006)[b]	8	100% L444P/L444P	N/C	As above with depression
Ida et al (1999)	15	100% L444P/L444P	17.1±11.4	Bulbar palsy Convulsion Hearing disability Horizontal gaze palsy/oculomotor apraxia Laryngospasms Mental retardation/low IQ Strabismus
Goldblatt et al (2005)	8	75% L444P/L444P	N/C	Oculomotor apraxia
Park et al (2003)	16	6.25% L444P/L444P	17.6 [3–40]	Cognitive deficits, delayed development Cerebellar findings Epilepsy Dysarthria Dystonia Halting or slowing of speech Swallowing difficulties Tremor Unsteady gait
Pasmanik-Chor et al (1996)[a]	12	100% D409H mutation	10.8 [2–20]	Oculomotor apraxia
Tylki-Szymanska et al (2006)	22	72% L444P/L444P	11.7	Epilepsy Mask-like expression Myoclonus retroflexion neck Strabismus Supranuclear gaze paresis Kyphosis
GOSH Single centre initial cohort	15	80% L444P/L444P	9 [2–20]	Ataxia Brisk deep tendon reflexes Cognitive impairment Dysphagia Dysarthria Horizontal gaze palsy/saccade initiation failure Intention tremor Pyramidal tract dysfunction Seizures Sixth-nerve palsy (convergent squint)
Pilot cohort	47	71.4% L444P/L444P	Median 13.9	

N/C, not clear.
[a] Duplication of the same cohort
[b] Follow-up of original cohort.

epilepsy as a stand alone term (Altarescu et al 2001; Erikson et al 1995, 2006; Ida et al 1999; Park et al 2003; Tylki-Szymanska et al 2006).

Cognitive ability of varying ability was reported in four of the cohorts. Again the classification and definitions reported varied, from mental retardation (IQ≤70) (Ida et al 1999), cognitive deficits, developmental delay and mental deterioration (Park et al 2003) and low cognitive function (Erikson et al 1995, 2006; Tylki-Szymanska et al 2006).

Ataxia or abnormal gait pattern was identified in two papers (Erikson et al 1995, 2006; Park et al 2003). These were the only papers to also note intention tremor and tremor.

Spasticity and dystonia were reported only once in Erikson and colleagues (1995) and Park and colleagues (2003) respectively, along with nonspecific cerebellar findings in the latter. Swallowing difficulties, bulbar palsy and laryngospasm were noted in two (Ida et al 1999; Park et al 2003) with dysarthria and halting or slowing of speech in only one (Park et al 2003). Strabismus appeared in two (Ida et al 1999; Tylki-Szymanska et al 2006). Erikson and colleagues (1995) and Tylki-Szymanska and colleagues (2006) were the only papers to describe kyphosis in their cohort.

Other findings identified in the papers are mask-like expression, retroflexion of neck and hearing difficulties. Whether these have a neurological pathology, however, is unclear. Dementia was only reported by Erikson and colleagues (1995, 2006) as in the original report of this cohort. Depression was identified in two papers (Erikson et al 2006; Park et al 2003). However, depression is a common feature of chronic disease in general and not exclusive to NGD.

This collection of publications presents accounts of the neurological features of 102 patients with NGD. Overall these findings are consistent with those of the internal cohort (Table 2). Although it is difficult to extrapolate data to meaningfully compare cohorts in terms of age, genotype, spleen status and ERT dose, it would appear that the presenting neurological features are very similar. The main difference is that progressive myoclonic epilepsy (n=2) and laryngospasms (n=3) are reported, although in very small numbers compared to the group as a whole. Therefore, there appeared to be no significant indication to change and/or amend the original features identified, particularly as the epilepsy and swallowing difficulties/oral bulbar domains capture these features.

Development of tool

Definitions of ‘mild’, ‘moderate’, ‘severe’ are at best subjective and therefore were not encouraged as stand-alone terms when creating variables to populate the domains of the SST. Categories of severity were defined for each of the original identified features, and weighted with equal scores from 0 to 3, with 3 indicating a more severe presentation of that feature and a total score of 39 indicating the most severe disease overall (Table 3).

Seizure management across centres is varied. Not all centres regularly monitor seizure activity with EEG. Some centres have an aggressive approach when introducing therapy, while others monitor isolated seizures and myoclonus seizures for some time before introducing therapy. Owing to the complexity of seizure management, a proposed criterion was identified. This would avoid the added complexity of involving EEG assessments and interpretation to grade severity.

Cognitive ability is another domain that is measured differently across centres. There are well-validated assessment tools to assess this, but they are time consuming, involve other specialist and have cultural variation. IQ ranges based on age appropriate assessment are included as indicators for assessment but should not be seen as an absolute for scoring. School achievements and expert interpretation compared to normal developmental milestones can be considered.

The domains that posed greatest difficulty in allocating grading were fine motor skills and gross motor skills as there are well-established assessments available to assess these functions, but they are often seen as laborious and time consuming. Selection of suitable assessment is also age dependent. In order to introduce a user-friendly scale, the decision was made to classify function as age appropriate/normal or not age appropriate/impaired.

Age is a significant challenge when assessing neurology. The sensitivity of assessment required when dealing with infants and toddlers is very difficult to capture and classify. This was clearly evident when populating the cognitive ability, ataxia/gait and speech domains, as highlighted later in the discussion. Some of these obstacles remain a challenge, which future work will aim to address.

Domain classification was then discussed with two paediatric neurologists and four European Working Group of Gaucher Disease (EWGGD) task force experts, and small changes were made.

Table 3 Version 1 of the severity scoring tool (SST)

Horizontal gaze palsy	Normal (although not likely in diagnosis)	0
	Horizontal saccades absent, vertical saccades present	1.5
	Horizontal saccades and vertical saccades absent	3
Epilepsy	No seizures	0
	Seizures not requiring anticonvulsants	1
	Seizures controlled with anticonvulsants	2
	Seizures requiring combination therapy or resistant to anticonvulsants	3
Development/cognitive ability	Normal	0
	Mildly impaired (IQ less than 85 or equivalent)	1
	Moderate (IQ between 50–57 or equivalent)	2
	Severe (more than half their chronological age)	3
Neurology pattern		
Ataxia/gait	Normal, apparent only on tandem walking	0
	Ataxia on straight gait, able to walk without assistance	1
	Able to walk only with assistance	2
	Unable to walk	3
Cerebellar signs/ataxia	No intention tremor	0
	Intention tremor not affecting function	1.5
	Intention tremor with marked impact on function	3
Pyramidal	Normal tone with increased reflexes	0
	Mildly to moderately increased tone and reflexes	1
	Increased tone reflexes with sustained/unsustained clonus	2
	Severe spasticity with inability to walk	3
Extrapyramidal	Normal	0
	Variable tone and posturing not impairing function, with or without therapy	1
	Variable tone and posturing impairing function, despite therapy	2
	Significant rigidity with no/minimal benefit from therapy	3
Swallowing difficulties/ oral bulbar function	Normal	0
	Mild dysphagia (excess drooling)	1
	Moderate dysphagia (risk of aspiration, modification to diet required)	2
	Severe dysphagia (requiring nonoral feeding)	3
Speech	Normal	0
	Mild to moderate dysarthria impairing intelligibility to unfamiliar listener	1
	Severe dysarthria with most speech unintelligible to familiar and unfamiliar listener	2
	Anarthria	3
Neurology function		
Fine motor skills	Age appropriate/Normal	0
	Not age appropriate/impaired	3
Gross motor skills	Age appropriate/Normal	0
	Not age appropriate/impaired	3
Ophthamology	Normal	0
	Cranial nerve palsy (previously corrected or not)	1.5
	Cranial nerve palsy (reappearing despite surgical correction)	3
Spinal alignment	Normal	0
	Mild kyphosis – but flexible	1
	Moderate kyphosis – partially corrected	2
	Severe kyphosis – fixed	3
	Other neurological features not captured	
Total		39

Severity scoring tool pilot use

With the collaboration and support of the EWGGD Task Force, 47 patients with NGD across four countries, were assessed using the scale. The countries involved were Sweden, Poland and Germany. In each instance the first author gave a detailed explanation of the tool to each clinician prior to examination of patients, and worked with each clinician in assessment.

Patients were assessed between November 2005 and February 2006. Oppenheim (2003) suggests a minimum of 100 respondents for an adequately informed pilot analysis. This was not possible in the present study given the rarity of NGD. The cohort had a large proportion of Polish and Norrbottnian ethnicity. The majority of the patients were homozygous for the L444P mutation at 75.6% (n=34), 4.4% L444P/D409H, 11.1% L444P/other, and a further 8.8% other. Forty-five patients were receiving ERT. Seventy-one per cent (n=32) had an intact spleen or partial splenectomy while 28.9% (n=13) had a complete splenectomy. Median age at assessment was 15.5 years (range 2.3–54.8 years). Median age at start of therapy was 12.6 years at a mean dose of 59 IU/kg per 2 weeks (SD±34.7). The mean dose of ERT at the time of assessment was 84.11 IU/kg per 2 weeks (SD±77.2). Two patients had undergone BMT.

The mean SST score on assessment was 8.044 (SD±5.58), range 1.5–22. This score differed across genotype, with L444P homozygotes scoring a slightly lower mean score of 7.3 (SD±4.36) and L444P/other scoring 12.3 (SD±10.2). Patients with an L444P/D409H genotype (n=2) had a mean score of 5.0 (SD±1.14), while all other genotypes scored 10.6 (SD±7.30).

Those who had undergone a complete splenectomy had a higher mean score of 10.0 (SD±4.63) compared to 7.234 (SD±5.79) in all others. The two BMT patients had a mean score of 10 (SD±2.0).

Internal reliability

The data set of 47 patients was put forward into the SPSS 'scale' window for reliability analysis. Initial exploration of correlation coefficient within the data using Spearman's rho demonstrated that ataxia/gait had the greatest number of significant correlations with other domains at the 5% level. Correlation coefficients are more discriminating the higher they are. A correlation of +1 indicates perfect agreement, with a minimum of 0.2 generally required (Oppenheim 2003). There was correlation above 0.2 for all the domains. There was high significant correlation with cerebellar signs/ataxia (r_s=0.298), pyramidal (r_s=0.644), and extrapyramidal (r_s=0.371), p=0.046, 0.000 and 0.012 significance, respectively. This statistical correlation is clinically expected as all four assess the neurology pattern. Whether the correlation is too high and measuring the same feature, however, needs to be considered. There was also high correlation with fine motor skills (r_s=0.398), and gross motor skills (r_s=0.495), p=0.007 and 0.001, respectively. All of these six domains correlated very highly with the SST overall.

A further four domains, extrapyramidal, speech, fine and gross motor skills, correlated significantly with seven other domains. Speech and swallowing difficulties also correlated well with each other (r_s=0.400) at a significance level of 0.006, which again has clinical relevance.

Horizontal gaze palsy only correlated significantly with two domains: ataxia/gait (r_s=0.394, p=0.007) and cerebellar signs/ataxia (r_s=0.372, p=0.012). It is uncertain whether any clinical conclusions can be drawn from this statistical correlation. Epilepsy correlated significantly with four other domains: ataxia/gait (r_s=0.297, p=0.048), extrapyramidal (r_s=0.337, p=0.024), speech (r_s=0.411, p=0.005), and gross motor skills (r_s=0.398, p=0.007).

It was interesting to note that cognitive ability did not statistically correlate with any of the other domains, with only three domains correlating above the 0.2 level. Indeed, cognitive ability did not correlate well with the SST overall (r_s=0.336, p=0.024). A possible clinical explanation for this is that the variable cognitive profile seen in these patients has been highlighted as variable, with performance and verbal IQ often scoring very differently (Durling et al 2006).

Similarly ophthalmology demonstrated relatively poor correlation, with three above the 0.2 level but none statistically significant. Of less surprise is that spinal alignment correlated significantly with only one domain, which was pyramidal (r_s=0.357, p=0.016).

Table 4 Reliability statistic (Cronbach's alpha) for each domain

Domain	Cronbach's alpha if item deleted
Horizontal gaze palsy	0.719
Epilepsy	0.748
Cognitive ability	0.740
Ataxia/gait	0.678
Cerebellar signs/ataxia	0.707
Pyramidal	0.658
Extrapyramidal	0.672
Swallowing difficulties	0.709
Speech	0.693
Fine motor skills	0.706
Gross motor skills	0.692
Ophthalmology	0.729
Spinal alignment	0.754

Table 5 Content validity criteria

1	Irrelevant item
2	Unable to assess relevance of item without item revision or item is in need of such revision that it would no longer be relevant
3	Relevant but needs minor alterations
4	Very relevant and succinct

Retaining or revising domains needs to be considered on a clinical and statistical basis. As domains must not be overly redundant (highly correlated with one another) a correlation coefficient >0.7 is generally accepted: any higher and eliminating a domain should be considered, without negatively impacting on the overall tool (Ruperto and Giannini 1996). In this instance the highest correlation coefficient was seen between ataxia/gait and pyramidal (r_s=0.644, p=0.000). None of the domains consistently correlated below 0.2.Therefore, none of the domains appears to be redundant.

The SST with 13 domains scored a Cronbach's alpha of 0.726, which is acceptable. The analysis demonstrates what Cronbach's alpha would be achieved if individual domains were removed from the tool (Table 4). Cognitive ability, ophthalmology and spinal alignment were the most poorly correlated domains. Removing ophthalmology, cognitive ability or spinal alignment individually improved the Cronbach's alpha by 0.003, 0.014 and 0.028, respectively. These equate to minimal improvement.

The lack of correlation of cognitive ability with other domains may simply be a reflection of the known discrepancy between verbal and nonverbal intelligence. Therefore, this domain was retained. There is no good evidence currently that kyphosis represents a neurological rather than a bone presentation, and it highlights an area for further clarification. Owing to the large number of patients in this cohort presenting with kyphosis (n=12), however, it was deemed a valuable to capture data until a better understanding of the pathology is achieved. The same argument is proposed for ophthalmology. The Cronbach's alpha has the advantage that testing reliability in the pilot analysis need only be done once to test responses. The minimal change in Cronbach's alpha and the clinical indications for these three domains are maintained in the SST.

Content validity

Determination of validity was undertaken following the steps outlined by Lynn (1986): establishing the assertion by experts that the individual domains within the instrument is valid and that the SST as an entire instrument is also valid.

Four experts were identified through the EWGGD Task Force, all authors of the 2001 published guidelines in the management of NGD (Vellodi et al 2001) with a fifth independent paediatric neurologist from a UK specialist centre. All were paediatric consultants with at least 10 years' experience with a specialist interest in NGD and all were involved in SST pilot use. The experts were provided with detailed information regarding the process and their role in evaluating the SST, as highlighted by Lynn (1986) (Table 4). They were requested to examine the SST and respond to the questionnaire, indicating their expert opinion on each neurological feature. The author was particularly

Table 6 Responders index of content validity (CVI) for each domain

Neurological domain	Number of responders scoring 3	Number of responders scoring 4	Actual CVI
Horizontal gaze palsy	1	4	1.00
Epilepsy	1	4	1.00
Development/cognitive ability	1	4	1.00
Ataxia/gait	2	3	1.00
Cerebellar signs/ataxia	1	4	1.00
Pyramidal	1	4	1.00
Extrapyramidal	0	4	0.80
Swallowing difficulties/oral bulbar function	2	3	1.00
Speech	2	1	0.60
Fine motor skills	1	2	0.60
Gross motor skills	1	1	0.40
Ophthalmology	4	1	1.00
Spinal alignment	2	2	0.80
SST overall	1	4	1.00

interested in whether each domain missed an aspect of that neurological domain, or was unclear. The experts, in addition to judging each domain, were asked to identify any neurological feature(s) that they felt had been omitted in error from the SST.

Communication was undertaken initially through e-mail and then personal contact. All comments were received within 2 weeks of initial correspondence and within 8 months of using the instrument. Responses were received from all five experts. Based on the number of expert responders, an agreement by four was required to establish a index of content validity (CVI) score of 0.80 or complete agreement to achieve a CVI of 1.00. The CVI is seen in Table 5.

Domain content validity

Eight out of the 13 domains revealed excellent CVI and established content validity beyond the 0.05 level of significance (Table 6). A further two domains achieved a CVI of 0.80. The neurological domains that demonstrated low validity were speech, fine motor skills and gross motor skills. General comments regarding 'speech' was that it was 'very subjective' and that it was 'inappropriate for the very young child'.

A smaller number of measures that yield the same clinical information with less complexity and a decrease in the probability of statistical error and ambiguous results are sought. Sensitivity of the tool is paramount. The crude nature of the fine motor skills and gross motor skills, by the nature of their presentation, fails to capture the same sensitivity as the other domains. Removal of these domains was therefore proposed to the group, and was agreed upon.

In response to the comments regarding the speech domain, the literature was reviewed again. Steinfeld and colleagues (2002) captured language in very similar format: Normal (3), Has become recognizably abnormal (2), Hardly understandable (1), and Unintelligible or no language (0). Another recently published scale for the assessment and rating of ataxia (SARA) (Schmitz-Hubsch et al 2006) offers a similar speech disturbance scoring category although with more variables: Normal (0) Suggestion of speech disturbance (1), Impaired speech but easy to understand (2), Occasional words difficult to understand (3), Many words difficult to understand (4), Only single words understandable (5), Speech unintelligible/anarthia (6). These are also subjective in nature, and do not offer an improved solution to the one currently proposed. In response to concerns about monitoring very young children, a bracketed comment was added to advise assessors which score to choose, e.g. (Normal (and those too young yet to speak)).

Instrument content validity

Four out of the five responders rated the SST overall as 'very relevant and succinct', while the fifth rated it as 'relevant but needs minor alterations', providing an index of content validity (CVI) for the entire SST instrument of 1.00.

Concurrent validity

As specified, there is currently no gold standard of measurement for NGD that could be used as a comparative assessment to demonstrate concurrent validity of this tool. An attempt was therefore made to identify whether scores reflected severity as would be expected based on the clinical presentation, and what could be regarded as genotype–phenotype correlation in this vast heterogeneous group. As demonstrated, the mean SST differed across genotype. L444P homozygotes scored lower at 7.3 (SD±4.36) with L444P/others scoring 12.3 (SD±10.2). This large standard deviation is most likely to be a representative of the great heterogeneity in this group, although it could also be argued to be a reflection of a relatively small sample size (n=5). Those without an L444P allele (n=4) scored 10.6 (SD±7.30). Patients with an L444P/D409H genotype (n=2) had a mean score of 5.0 (SD±1.14). The elder of the two BMT patients scored the highest, which would be expected in a progressive disease (mean score of 10 (SD±2.0)).

As expected, those who had undergone a complete splenectomy had a higher mean score of 10.0 (SD±4.63) compared to 7.234 (SD±5.79). This indicates concordance with the expected increased severity in this group.

Feasibility

Although not systematically evaluated, the feasibility of using the SST was evident. Assessments were achieved easily and quickly within a 20–30-minute clinical examination and discussion with patient and family. The assessing physician did not need to rely on any other specialist to make assessment, and although previous IQ assessments offered an exact evaluation of cognitive ability this was not an absolute necessity to enable assessment. Patients were exposed to minimal or no discomfort, and there were no risks during examination. Importantly, the assessment did not involve additional cost.

Table 7 Final version of NGD severity scoring tool[a]

Horizontal gaze palsy	Normal (although not likely in diagnosis)	0
	Horizontal saccades absent, vertical saccades present	1.5
	Horizontal saccades and vertical saccades absent	3
Epilepsy	No seizures	0
	Seizures not requiring anticonvulsants	1
	Seizures controlled with anticonvulsants	2
	Seizures requiring combination therapy or resistant to anticonvulsants	3
Development/ cognitive ability	Normal	0
	Mildly impaired (IQ less than 85 or equivalent)	1
	Moderate (IQ between 50–57 or equivalent)	2
	Severe (more than half their chronological age)	3
Neurology pattern		
Ataxia/ gait	Normal, apparent only on tandem walking	0
	Ataxia on straight gait, able to walk without assistance	1
	Able to walk only with assistance	2
	Unable to walk	3
Cerebellar signs/ataxia	No intention tremor	0
	Intention tremor not affecting function	1.5
	Intention tremor with marked impact on function	3
Pyramidal	Normal tone with increased reflexes	0
	Mildly to moderately increased tone and reflexes	1
	Increased tone reflexes with sustained/unsustained clonus	2
	Severe spasticity with inability to walk	3
Extrapyramidal	Normal	0
	Variable tone and posturing not impairing function, with or without therapy	1
	Variable tone and posturing impairing function, despite therapy	2
	Significant rigidity with no/minimal benefit from therapy	3
Swallowing difficulties/oral bulbar function	Normal	0
	Mild dysphagia (excess drooling)	1
	Moderate dysphagia (risk of aspiration, modification to diet required)	2
	Severe dysphagia (requiring nonoral feeding)	3
Speech	Normal (and those too young yet to speak)	0
	Mild to moderate dysarthria impairing intelligibility to unfamiliar listener	1
	Severe dysarthria with most speech unintelligible to familiar and unfamiliar listener	2
	Anarthria	3
Ophthamology	Normal	0
	Cranial nerve palsy (previously corrected or not)	1.5
	Cranial nerve palsy (reappearing despite surgical correction)	3
Spinal alignment (kyphosis)	Normal	0
	Mild kyphosis – but flexible	1
	Moderate kyphosis – partially corrected	2
	Severe kyphosis – fixed	3
	Other neurological features not captured	
Total		33

[a] NGD severity scoring tool developed at Great Ormond Street Hospital for Children NHS Trust (2006). © Copyright GOSH (2006).

Discussion

This study represented the largest cohort of NGD patients ever assessed uniformly. The final SST (Table 7) is proposed as a concise, user-friendly and systematic tool to evaluate the neurological manifestation of NGD patients. The demonstrated validity makes it a suitable tool for clinical practice.

The clinical value of retaining cognitive ability, ophthalmology and spinal alignment, considering their relatively poor statistical correlation with the other domains and with the SST overall, highlighted an area for discussion; and an argument was made for retaining them, which was supported by the Cronbach alpha, clinical findings and content validity assessment.

Although an index of content validity (CVI) of 1.00 for the entire SST instrument was scored, owing to the poor individual domain response to fine motor skills and gross motor skills, and their lack of sensitivity, they were removed, now offering 11 domains.

Reproducibility assessment of the SST is currently under way, examining intra-rater and inter-rater reliability, where repeat assessments of patients takes place at 3-monthly interval as disease progression would not be expected in this time frame. This assessment will also examine discriminant validity, and the SST's sensitivity to demonstrate change in longitudinal observation in conjunction with a retrospective analysis of patients' notes.

A limitation of the tool is that it fails to incorporate the impact of systemic and visceral disease upon neurological function. This needs to be explored further, possibly in parallel with other tools. A significant limitation of the SST is in the assessment of young children. Further work will need to done to incorporate developmental quotient ranges for the cognitive ability domain, allowing children younger than 4 years to be assessed uniformly. Ataxia is another domain that needs to be classified for younger children who are not of walking age. Although the SST was used to assess two children under 5 years old, it clearly has limited sensitivity in this age group currently. More work will now be done in this field.

A further development of the tool currently under way is to explore the burden of each individual domain upon disease severity. It is anticipated, for example, that epilepsy has a greater severity and burden of disease than ophthalmology, and will therefore need to be 'weighted' with a higher score. This will be done through a combination of statistical analysis and seeking the opinion of experts and patients and their families. Additionally, a consensus among expert clinicians on the minimal clinically important difference scale score will need to be identified. This is the smallest difference in score which clinicians perceive as improvement or deterioration in disease.

Conclusions

A retrospective analysis of 15 patients and a literature review of 102 patients have provided the basis for this tool. Pilot analysis of its use on a cohort of 47 patients has enabled internal reliability and validity to be demonstrated. Five European specialists, all experts in the management of NGD patients, have objectively evaluated the content validity. The revised SST offers a feasible measurement scale of NGD. It is easy and relatively quick to apply, has no cultural or economic constraints, and can be used at no additional cost. The revised instrument (Table 7) is available from the lead author.

Acknowledgements This work was supported by a travel grant from the Gaucher Association. We would like to thank all the experts who participated in the study: Dr Anna Tylki-Szymanska (Department of Metabolic Diseases, The Children's Memorial Health Institute, Warsaw, Poland), Dr Eugen Megel (Mainz) and Dr Anders Erikson (Department of Paediatrics, Umea University, Umea, Sweden).

References

Abrahamov A, Elstein D, Gross-Tsur V, et al (1995) Gaucher's disease variant characterised by progressive calcification of heart valve and unique genotype. *The Lancet* **346**: 1000–1003.

Altarescu G, Hill S, Wiggs E, et al (2001) The efficacy of enzyme replacement therapy in patients with chronic neuronopathic Gaucher's disease. *J Pediatr* **138**(4): 539–547.

Bowling, A (2001) *Measuring Disease*, 2nd edn. Milton Keynes: Open University Press.

DeVellis RF. (1991) *Scale Development: Theory and Applications.* Newbry Park, CA: Sage Publications.

Durling E, Dale N, Davies E, Vellodi A (2006) Understanding the cognitive profile of children with neuronopathic Gaucher disease. Poster presentation. European Working Group on Gaucher Disease, Cambridge July 2006.

Erikson A, Dreborg S, Hagberg B (1980) Gaucher disease – Norrbottnian type. General clinical description. *Eur J Pediatr* **133**(2): 107–118.

Erikson A, Astrom M, Mansson JE (1995) Enzyme infusion therapy of the Norrbottnian (Type 3) Gaucher disease. *Neuropediatrics* **26**(4): 203–207.

Erikson A, Forsberg H, Nilsson M, et al (2006) Ten years' experience of enzyme infusion therapy of Norrbottnian (type 3) Gaucher disease. *Acta Paediatr* **95**(3): 312–317.

Gibson F, Cargill J, Allison J, et al (2006) Establishing content validity of the oral assessment guide in children and young people. *Eur J Cancer* **42**(12): 1–9.

Goldblatt J, Szer JM, Fletcher J, et al (2005) Enzyme replacement therapy for Gaucher disease in Australia. *Int Med J* **35**: 156–161.

Ida H, Rennert OM, Iwasawa K, Kob (1999) Clinical and genetic studies of Japanese homozygotes for the Gaucher disease L444P mutation. *Hum Genet* **105**(1–2): 120–126.

Lynn MR (1986) Determination and quantification of content validity. *Res Nurs Health* **35**(6): 382–385.

Oppenheim A (2003) *Questionnaire, Design, Interviewing and Attitude Measurement*. New York: Continuum.

Park JK, Orvisky E, Tayebi N, et al (2003) Myoclonic epilepsy in Gaucher disease: genotype–phenotype insights from a rare patients subgroup. *Pediatr Res* **53**(3): 387–395.

Pasmanik-Chor M, Laadan S, Elroy-Stein O, et al (1996) The glucocerebrosidase D409H mutation in Gaucher disease. *Biochem Mol Med* **59**(2): 125–133.

Richmond JP, Wright ME (2005) Development of a constipation risk assessment scale. *Clin Effect Nurs* **9**: 37–48.

Ruperto N, Giannini EH (1996) Redundancy of conventional articular response variables used in juvenile chronic arthritis clinical trials. *Ann Rheum Dis* **55**(1): 73–75.

Rusca CL (1921) Sul morbo del Gaucher. *Hematologica* **2**: 441.

Rust J, Golombok S (1995) *Modern Psychometrics. The Science of Psychological Assessment*. London: Routledge.

Seong TJ (2002) *Validity and Reliability*. Seoul: Hakjsa.

Schmitz-Hubsch T, du Montcel ST, Baliko L, et al (2006) Scale for the assessment and rating of ataxia: development of a new clinical scale. *Neurology* **13**: 1717–1720.

Steinfield R, Heim P, von Gregory H, et al (2002) Late infantile neuronal ceroid lipofuscinosis: quantitative description of the clinical course in patients with CLN2 mutations. *Am J Med Genet* **112**: 347–354.

Tylki-Szymanska A, Keddache MS, Grabowski GA (2006) Characterization of neuronopathic Gaucher disease among ethnic Poles. *Genet Med* **8**(1): 8–15.

van den Hout HM, Hop W, van Diggelen OP (2003) The natural course of infantile Pompe's disease: 20 orignial cases compared with 133 cases from the literature. *Pediatrics* **112**: 332–340.

Vellodi A, Bembi B, de Villemeur TB, et al (2001) Management of neuronopathic Gaucher disease: a European consensus. *J Inherit Metab Dis* **24**: 319–327.

Whybra C, Kampmann C, Krummeneauer F et al (2004) The Mainz Severity Index: a new instrument for quantifying the Anderson-Fabry disease phenotype and response of patients to enzyme replacement therapy. *Clin Genet* **65**: 299–307.

Woolery M, Carroll E, Fenn E, et al (2006) A constipation assessment scale for use in pediatric oncology. *Pediatr Oncol Nurs* **23**(2): 65–74.

Zimran A, Sorge J, Gross E, Kubitz M, West C, Beutler E (1989) Prediction of severity of Gaucher's disease by identification of mutations at DNA level. *Lancet.* **2**: 349–352.