A validated disease severity scoring system for adults with type 1 Gaucher disease

Neal J. Weinreb, MD¹, Maria D. Cappellini, MD², Timothy M. Cox, MD³,

Edward H. Giannini, MSc, DrPH⁴, Gregory A. Grabowski, MD⁵, Wuh-Liang Hwu, MD, PhD⁶, Henry Mankin, MD⁷, Ana Maria Martins, MD⁸, Carolyn Sawyer, PhD⁹, Stephan vom Dahl, MD¹⁰, Michael S. Yeh, MD, MPH⁹, and Ari Zimran, MD¹¹

Purpose: A validated disease severity scoring system (DS3) for Gaucher disease type 1 (GD1) is needed to standardize patient monitoring and to define patient cohorts in clinical studies. **Methods:** DS3 domains were established by an expert physician group using the nominal group technique of consensus formation. Items were selected by 36 GD1 physicians. The expert group determined appropriate measurement techniques for each item. Measurements were weighted considering contributions to GD1 morbidity and mortality. Consensus Clinical

Carolyn Sawyer is currently at Novartis Biomedical Research Institute, Cambridge, MA.

Neal J. Weinreb, MD, University Research Foundation for Lysosomal Storage Disorders, 8170 Royal Palm Boulevard, Coral Springs, FL 33065. E-mail: boneal@winning.com.

Disclosure: This project was sponsored by Genzyme. NJW reports receiving honoraria from Genzyme, Amicus, and Shire, receiving travel grants from Amicus, receiving an educational grant from Genzyme, serving as an investigator on clinical trials for Shire, Amicus, Protalix, and Genzyme, and serving as on advisory boards for Genzyme, Shire, and Protalix. MDC reports receiving honoraria from and serving on advisory boards for Novartis and Genzyme. TMC reports receiving honoraria from Genzyme, Shire HGT, and Actelion, serving as an investigator on clinical trials for Genzyme, Amicus, Actelion, and Protalix, and receiving unrestricted laboratory research support from Shire HGT and Genzyme. EG has served as a consultant for Genzyme. GAG reports receiving honoraria from Genzyme and Shire HGT, receiving research contracts from Genzyme, Shire HGT and Amicus, serving on the board of Project Hope and the Genzyme Gaucher Initiative, and serving as a consultant for Genzyme, Shire HGT, and Amicus. W-LH reports receiving honoraria and research contracts from Genzyme. HM has served as a consultant for Genzyme. AMM reports receiving honoraria from Genzyme, travel grants and research contracts from Genzyme and Shire, serving on advisory boards for Genzyme and Shire and serving as a consultant for Genzyme. CS and MSY were full-time employees of Genzyme at the time this work was performed and both hold stock in Genzyme. SvD reports serving as a consultant for Genzyme and Actelion, receiving honoraria from Genzyme, Actelion, and Shire, serving on advisory boards for Genzyme and Shire, and receiving research contracts from Genzyme. AZ reports serving as a consultant for Shire HGT and Protalix, receiving honoraria from Actelion, Genyzme, and Shire, serving on an advisory board for Protalix, and receiving stock options from Protalix.

Submitted for publication May 22, 2009.

Accepted for publication September 29, 2009.

Published online ahead of print December 18, 2009.

DOI: 10.1097/GIM.0b013e3181c39194

Global Impression Severity scores for sample cases were compared with average DS3 scores. A minimal clinically important difference in GD1 DS3 score was calculated. **Results:** The GD1 DS3 includes bone (42% of score), hematologic (32%), and visceral domains (26%); individual items use routine assessments, including medical history, blood chemistry, organ volume measurements, and bone evaluations (magnetic resonance imaging and dual x-ray absorptiometry). The maximum score is 19. Interrater reliability was 0.97 (Cohen's kappa). DS3 scores were highly correlated with Clinical Global Impression Severity scores ($r^2 =$ 0.89). The minimal clinically important difference was -3.2 improvement and +3.9 deterioration. **Conclusion:** This DS3 accurately quantifies GD1 status and intrapatient change over time. Testing of reliability and validity will continue to allow eventual implementation of the DS3 in clinical studies and routine practice. **Genet Med 2010:12(1):44–51.**

Key Words: Gaucher disease, severity scoring, inherited metabolic disorders, DS3, acid β -glucosidase

aucher disease (GD) is an autosomal-recessive inborn error Jof glycosphingolipid metabolism caused by loss-of-function mutations in the gene encoding lysosomal glucocerebrosidase (GCase; EC 3.2.1.45). When GCase activity is diminished, glucosylceramide accumulates primarily in the lysosomes of macrophages ("Gaucher cells"). GD is broadly subdivided into neuronopathic (types 2 and 3) and non-neuronopathic (type 1) phenotypes. GD type 1 (GD1, OMIM 230800), the most common variant with an estimated prevalence of between 1:40,000 and 1:60,000,1-3 is traditionally defined by the absence of primary central nervous system involvement. Worldwide, GD1 constitutes 95% of all cases,⁴ although the relative number of patients with neuronopathic variants is often substantially greater in Asian, Middle Eastern, and African populations. The common clinical manifestations are hematological cytopenias, hepatomegaly, splenomegaly, and various skeletal disorders. Disease expression is diverse. Some individuals homozygous for mutations in the glucocerebrosidase gene, especially those of Ashkenazi Jewish origin, are asymptomatic or are only mildly affected, at least until late in adulthood.4,5 However, others, including Ashkenazi Jews, have more severe and potentially disabling manifestations including irreversible bone disease with osteolytic lesions, avascular necrosis and marked osteoporosis.5,6 The rate and extent of disease progression are variable. Treatment, most commonly with enzyme replacement therapy, effectively ameliorates many of the manifestations of GD1.7-10

However, because genotype-phenotype correlations are imprecise, it is difficult to predict who will benefit from expensive, potentially life-long therapies and who may suffer if treatment is withheld, interrupted, or otherwise modified. A validated, reproducible, and broadly applicable tool to classify GD1 severity at any stage of the disease would greatly facilitate clinical studies that seek to identify predictive factors for prognosis or define disease management criteria.

Genetics IN Medicine • Volume 12, Number 1, January 2010

Copyright @ American College of Medical Genetics. Unauthorized reproduction of this article is prohibited

From the ¹University Research Foundation for Lysosomal Storage Disorders, Coral Springs, FL; ²Department of Internal Medicine, University of Milan, Milan, Italy; ³Department of Medicine, Addenbrooke's Hospital, University of Cambridge, United Kingdom; ⁴Division of Rheumatology and ⁵Division of Human Genetics, Cincinnati Children's Hospital Medical Center, Cincinnati, OH; ⁶Department of Pediatrics and Medical Genetics, National Taiwan University Hospital, Taipei, Taiwan; ⁷Department of Orthopedic Surgery, Massachusetts General Hospital, Boston, MA; ⁸Department of Pediatrics, Universidade Federal de Sao Paulo, Sao Paulo, Brazil; ⁹Genzyme Corporation, Cambridge, MA; ¹⁰Klinik für Innere Medizin, St. Franziskus-Hospital, Köln, Germany; and ¹¹Sha'are Zedek Medical Center, Jerusalem, Israel.

Current severity scoring indices for GD, such as the Zimran Severity Score Index (SSI)11 and Hermann score for bone disease,¹² were conceived in the pretreatment era. These systems include quantifiable parameters that can change over time or with treatment (organ volumes, serum liver-related tests, subjective pain assessments) but also assign substantial weight to constants (splenectomy, osteonecrosis, imaging abnormalities) and thus may not be sufficiently sensitive to define slowly responding therapeutic changes.13 The Zimran SSI score has not been tested with contemporary methods to assess performance parameters nor has its responsiveness to intrapatient changes in disease manifestations over time been established. As a result, the instrument has not been widely used in routine clinical practice or as an endpoint in clinical trials. More recently, an alternate SSI (GauSSI-I) has been proposed for GD1.14 However, this scoring system has yet to be validated and its routine use may be limited because of its complexity and the technological requirements of some of the component assessments. It may prove useful in limited applications such as single-center clinical trials.

A disease severity scoring system (DS3) expresses an integrated assessment of the burden of disease in a given patient. Ideally, the DS3 could be used to assess patient status, determine endpoints in clinical studies, classify disease subgroups, and compare outcomes among patients with similar levels of disease severity. DS3s use a minimal data set to score the patient in a comprehensive manner using groups of domains (often by organ system) that are populated with nonredundant items. Items should be valid and reliable, use feasible, standardized methods of assessment, and be weighted based on associated morbidity and mortality. A validated DS3 for GD1 is needed to guide clinicians as to when GD-specific therapy should be started, to monitor disease progression and treatment response in individuals, and to compare patient cohorts in clinical studies. Such disease severity scoring systems have already been developed for other chronic diseases such as rheumatoid arthritis.^{15,16} Here, we describe the development of a DS3 for adult GD1 patients as well as testing of the instrument for validity (content, face, criterion, discriminant, and construct validity), reliability, and feasibility.

MATERIALS AND METHODS

Instrument development

Domain selection

Recognized experts in GD1 were invited to participate and a DS3 Working Group comprised of nine GD1 experts from across the globe, representing multiple medical specialties, was formed; a methodologist-biostatistician experienced in the development of instruments for outcomes measurement was engaged to facilitate construction and validation of the GD1-DS3. A consensus conference was held in December 2006 to initiate development of the instrument. Putative domains were identified using the nominal group technique (NGT) of consensus formation.¹⁷

Domains and items

Seventy-four international non-Working Group physicians with experience in GD were invited in an electronic survey to propose quantifiable items for each domain that they considered "most relevant to the assessment and quantitative measurement of the activity/severity of Gaucher disease." They were also asked to rank the proposed domains in order of importance and to propose additional domains. The Working Group then used NGT to determine whether any domains originally proposed should be eliminated or whether new domains should be included. Items proposed were evaluated in terms of reliability, feasibility, and validity; duplicative or unnecessary items were eliminated. The methods of assessment appropriate for each item were selected by the Working Group using NGT consensus. A preliminary weighting system for items and domains was arrived at by NGT.

Validation and optimization

Content, face, criterion, discriminant, and construct validity were evaluated at two face-to-face conferences attended by 12 international experts in GD with no prior exposure to the DS3 (Table 1).¹⁸ Validation conference participation and logistics were arranged to coincide with major US and European Gaucher-related medical congresses. Deidentified patient profiles (n = 20) were suggested by Working Group members and were obtained from the International Collaborative Gaucher Group Gaucher Registry. Each patient profile contained complete clinical and diagnostic information from two sequential clinic visits, an initial and a follow-up visit, and was selected regardless of treatment status (all patients were untreated at the time of the initial visit). No values were imputed. Twenty profiles were chosen to achieve a reasonable assessment of sensitivity and specificity while limiting the time investment involved in scoring. Patient profiles were scored using the Clinical Global Impression (CGI) scale, which is used in a variety of other disease states.^{19–21} Each profile was scored for each visit for the overall extent of disease severity summarized as the CGI-Severity (CGI-S) score. The degree of change between visits was characterized using a CGI-Improvement score to rate each patient's status at the follow-up visit as "improved," "not changed," or "worsened" from the initial visit. Participants were also asked to determine whether the change in the patient's status would trigger a change in treatment or prognosis. NGT was used to achieve consensus on CGI scores and therapy/ prognosis change. Participants then individually scored the patient profile for each visit using the DS3. Results from the two individual conferences were not substantially or statistically different and were combined for analyses.

Participants also rated the feasibility and content validity of the DS3 using a validated technique that uses a 4-point scale, where 1 represented the worst feasibility/content validity and 4 represented the best feasibility/content.²² Additional feasibility and content validity testing was later carried out when an additional group of 23 physicians attending a GD workshop reviewed the DS3, scored a sample case, and rated the DS3 for feasibility and content validity using the same scale used at the validation conferences. Results from all 35 ratings were used to generate Feasibility Index and Content Validity Index scores (see Statistical Methods section).

An additional correlation exercise compared CGI-S scores with that of the Zimran SSI.¹¹ The original Working Group scored the 20 GD1 patient profiles for the initial and the follow-up visits using the Zimran SSI. For each patient, Zimran SSI scores were compared with the CGI-S scores obtained at the second validation conference.

Inclusion of items was refined according to statistical testing of reliability and stability. Additionally, scaling of and maximum scores for individual assessments within the DS3 were optimized to maximize the correlation between the total scores of the instrument and the consensus CGI-S scores using Excel 2002 Solver; this tool uses the Generalized Reduced Gradient-2 algorithm for optimizing nonlinear problems. The DS3 was optimized to maximize correlation with the CGI-S under conditions where bone density and infiltration data (the items judged most likely to be missing) were either present or absent.

Copyright Content American College of Medical Genetics. Unauthorized reproduction of this article is prohibited

,									
Feasibility		Easily performed in the routine clinic							
		nexpensive and quick							
		Low risk to the patient							
Reliability (reproducibility)		Of methods of assessments of each variable							
		Intra- and inter- rater							
		Of the instrument overall (internal consistency/reliability)							
Validity	Face	Clinical sensibility to the end-user							
	Content	Both consensus and quantitative							
		Criteria must consider all relevant aspects of disease that contribute to morbidity, without the inclusion of redundant items							
	Criterion	Agreement with a gold standard criterion							
		Sensitivity, specificity, and area under the curve							
		Predictive validity							
	Discriminant	Ability to pick up smallest amount of change considered to be the minimal clinica important difference (MCID)							
		Responsiveness to change over time							
	Construct	Criteria do what they are designed to do							
	Convergent	Results correlate with other tools that measure same construct							
	Divergent	Results do not correlate with tools that measure different constructs							

Table 1 Commonly used criteria for evaluating disease scoring systems¹⁸

Determination of the minimal clinically important difference

Minimal clinically important difference (MCID) represents the amount of change, either increase or decrease, in the instrument score that would usually indicate a change in some aspect of the disease or trigger an adjustment in medical care or prognosis. It is not synonymous with a difference that is statistically significant. We estimated MCID for the DS3 using data from the validation conference. Physicians were asked whether they felt that the observed change from one visit to the next would worsen, not change, or improve the patient's prognosis. Using CGI, the physicians came to a consensus decision regarding these two standards for each of the 20 sample patient profiles, given the change in the DS3. MCID was calculated as the amount of change in the DS3 score at which at least 75% of the physicians agree that there would be a change in prognosis.23 MCID for worsening was the change in the DS3 score at which at least 75% of physicians agreed that the patient's prognosis had worsened, and MCID for improvement was the change in the DS3 score at which at least 75% of physicians agreed that the patient's prognosis had improved.

Statistical methods

The Pearson rank coefficient was used to determine the correlation between the SSI and the DS3 scores. Each patient profile included data from an initial and follow-up visit, and each visit was scored using both the SSI and DS3. To increase the precision of the estimate of Pearson's rank coefficient (R), data from both the initial and follow-up visits were used in the calculation. Interrater reliability of the optimized DS3 instrument was assessed by having two independent raters score 10 of

the patient profiles, and the level of agreement was estimated using Cohen's kappa coefficient.

The Content Validity Index is the proportion of physicians who scored the instrument as either 3 (relevant but needs minor alterations) or 4 (very relevant and succinct). The Feasibility Index was calculated in the same manner as the Content Validity Index, with a score of 3 equal to feasible but needs minor alterations and a score of 4 equal to very feasible. A value of 0.8 or greater indicates that at least 80% of respondents find the instrument acceptable with respect to the quality assessed.

RESULTS

Instrument development

Domain and item selection

Six domains were initially selected for inclusion in the DS3: bone disease/skeletal, growth/metabolism, hematological, visceral, patient reported, and physician reported. Thirty-six of 74 (49%) respondents from around the world completed the electronic survey. Table 2 shows the relative importance of each domain as ranked by the survey respondents.

The three lowest-rated domains (patient reported, physician reported, and growth/metabolism) were eliminated. Patient-reported pain was integrated into the bone domain. The primary impact on the health-related quality of life in patients with GD1 is pain.²³ Including the pain component in the bone domain captures this dimension with less risk of redundancy. Fatigue, another patient-reported item, was considered by the Working Group to be a symptom often associated with other disease manifestations (e.g., anemia) and was therefore not incorporated. The growth/metabolism domain was eliminated, because

Domain	1	2	3	4	5	6	Average ran	
Bone disease/skeletal	17 (53%)	4 (13%)	5 (16%)	1 (3%)	0 (0%)	5 (16%)	2.31	
Growth/metabolism	3 (9%)	3 (9%)	5 (16%)	6 (19%)	12 (38%)	3 (9%)	2.76	
Hematological	7 (21%)	13 (39%)	3 (9%)	4 (12%)	3 (9%)	3 (9%)	3.63	
Patient-reported	2 (6%)	5 (16%)	8 (25%)	3 (9%)	7 (22%)	7 (22%)	3.77	
Physician-reported	7 (23%)	3 (10%)	3 (10%)	5 (16%)	3 (10%)	10 (32%)	3.91	
Visceral	0 (0%)	7 (20%)	9 (26%)	11 (31%)	6 (17%)	2 (6%)	3.94	
Values indicate count and per $1 = most$ important and $6 =$	ercentage of respond least important.	ents selecting the op	ption.					

Table 2 Importance of each domain, as ranked by 36 Gaucher disease physicians in the initial survey

it applies primarily to children and adolescents, whereas this instrument is designed solely for adult patients. Finally, the physician-reported domain, including a physician's global assessment, was eliminated because in rare diseases in which caregivers often see too few patients to develop clinical expertise the physician's global assessment may not be reliable. All items in the hematologic domain (thrombocytopenia, bleeding, anemia) were retained, because each is considered to be a clinically significant, nonredundant manifestation of GD1.

Selection and weighting of item measurements

For each item, the Working Group evaluated all known assessment options and selected methods of measurement based on feasibility and availability of technology, with the intention of containing costs and avoiding additional risks to patients. Preliminary weighing of measurements arrived at by Working Group consensus was assigned according to the extent that morbidity and mortality in each measurement contributes to

Gaucher Disease Type 1 Severity Scoring System (GD-DS3)

Patient ID

Assessment Date

DISEASE	ASSESSMENTS	DISEASE SEVERITY SCORE											Average	
DOMAINS		0	1	2	3	4	5	6	7	8	9	10	Assessment Score	Domain Score
BONE	Lytic Lesions, AVN, or Pathological Fractures*	Absent								Present			8	
	Bone / Joint Pain (past 30 days)	None to very mild pain		Mild pain			Moderate pain			Severe pain		Extreme pain		
		$\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 $								(100) (100)		(1) (1) (1) (1) (1) (1) (1) (1) (1) (1)	10	8.0
	# Bone Crisis in Past 12 Months	0 to 1		≥2									2	
	Bone Marrow Infiltration**	0 to 4 (mild)								5 to 8 (moderate)		9 to 16 (marked to severe)	10	
	Bone Mineral Density Z-score [†]	> -1		> –2 to ≤ –1								≤-2	10	
HEMATOLOGIC	Thrombocytopenia	≥ 120 x10 ³ /mm ³		21 to 119 x10 ³ /mm ³			< 20 x10 ³ /mm ³						5	
	Bleeding	None to mild tendency; bruising		Moderate; no transfusions						Severe; transfusion needed			8	6.0
	Anemia	> 12 g/dL (male), > 11 g/dL (female)		8 to 12 g/dL (male), 8 to 11 g/dL (female)			< 8 g/dL						5	
VISCERAL	Splenomegaly [‡] (Volume as MN)	≤ 5		> 5 to ≤ 15			> 15 or Splenectomized						5	
	Hepatomegaly [‡] (Volume as MN)	≤ 2.5		> 2.5									2	5.0
	Gaucher-related Pulmonary Disease	None								Any			8	
										Tot		Gaucher I	DS3 Score	19.0

AVN: avascular necrosis; MN: Multiples of Normal

*New within the past 12 months

**Bone marrow infiltration as measured by MRI using the Bone Marrow Burden score²⁴ or other quantitative or qualitative assessment of disease burden as mild, moderate, or marked to severe. ¹Bone mineral density Z score to be measured by dual-energy X-ray absorptiometry (DXA) if available; computed tomography or other methods are acceptable as long as results are expressed as a Z score.

¹Splenomegaly and hepatomegaly to be measured by (1) MRI or CT, (2) ultrasound, or (3) physical exam, in this order of preference, depending on technology available.

Fig. 1. Gaucher Disease Type 1 Disease Severity Scoring System instrument.

Genetics IN Medicine • Volume 12, Number 1, January 2010

Gaucher Disease Type 1 Disease Severity Scoring System (GD1 DS3) Instructions for Use

- 1. Record the date of assessment.
- 2. For each assessment, determine the GD1 DS3 score of the patient at the time of evaluation. (See "Notes Regarding Specific Assessments" below.)
 - a. If current data are not available for all assessments at the time the DS3 score is calculated, data from previous evaluations may be used to calculate the current DS3 score if the patient's overall clinical status has remained stable and assessments were collected within the following period of time prior to the current date:
 - Bone imaging: 12-24 months^{**}
 - Hematological: 12-24 months^{**}
 - Visceral imaging: 12-24 months^{**}
 - b. If data are not available (NA) at the time of assessment or from previous evaluations, select "NA" from the drop down menu (or write "NA" on printed copy).
 - The GD1 DS3 score has been optimized to be accurate and consistent with or without bone mineral density and bone marrow infiltration scores.
 - All other assessment scores, within the time frames described above, are required.
- 3. The tool can be adapted in a spreadsheet to automatically calculate average Disease Domain Scores and the total DS3 Score. To calculate the DS3 score manually:
 - a. First calculate the average Disease Domain Scores by adding the assessment scores for each domain (bone, hematological, and visceral) and dividing by the number of assessment scores completed. Do not include assessments marked "NA".
 - b. The total GD1 DS3 score is the sum of the three Disease Domain Scores.
- 4. The maximum possible GD1 DS3 score is 19. In initial validation testing using 20 patient cases scored at 2 different time points, no patient received a score higher than 13 and scores above 9 correlated with an assessment of "severe" disease. In qualitative terms, DS3 scores may be interpreted as follows:
 - 0 3 Borderline to mild disease
 - 3 6 Moderate disease
 - 6 9 Marked disease
 - 9+ Severe disease
- 5. The DS3 instrument is most useful when a complete set of data is available as the absence of assessments will reduce the accuracy of the DS3 instrument. Regular and comprehensive monitoring of the recommended assessments is advised not only to obtain the most accurate measure of severity with the DS3, but also to allow modification of disease management based upon disease progression and/or therapeutic response. Perform assessments at the time of each recommended evaluation.⁷⁻¹⁰

Notes Regarding Specific Assessments

- 1. Lytic lesions, AVN, or pathologic fracture: "present" means any new occurrence(s) in the past 12 months
- Bone marrow infiltration assessment may be reported either quantitatively (BMB Score) or qualitatively (mild, moderate, marked to severe).
- For bleeding, an assessment of moderate (no transfusions) or severe (transfusion needed) should be based on bleeding considered by the assessor to be related to Gaucher disease, whether due directly to low platelet count or due to related hematologic abnormalities.
- 4. Pain-related assessments should be in the absence of analgesics.
- 5. When assigning a score for a particular assessment, consider only the effects of Gaucher disease. For example, the assessment "Bone / Joint Pain" must consider only pain resulting from Gaucher disease, not pain attributable to other factors such as injury, arthritis, etc.

* Range of dates for evaluations are based upon recommended schedule of follow-up examinations for patients with stable Gaucher disease.⁷⁴⁰

Disclaimer: The GD1 DS3 score is not intended to replace comprehensive clinical evaluations and cannot substitute for, but only assist with, clinical judgment.

Fig. 1. (Continued).

GD1 severity. Weighing was optimized as described to obtain the final working model of the GD1-DS3 (Fig. 1).

Working model and scoring method

The GD1-DS3 has three domains: bone, hematological, and visceral. Each domain contains three or more items, each scored individually by the evaluating physician. The domain score is tabulated by averaging the scores for all items within the domain. A total GD-DS3 score is the sum of the three domain

scores, with a maximum score of 19 points. A scoring reference guide was developed.

Validation

Feasibility and content validity

The Content Validity Index was 0.96, and the Feasibility Index was 0.95.

Copyright © American College of Medical Genetics. Unauthorized reproduction of this article is prohibited



Fig. 2. Correlations between the Clinical Global Impression Scale–Severity (CGI-S) and the Gaucher Disease Type 1 Disease Severity Scoring System (GD1-DS3) score for 20 patient profiles, each scored at two different time points by 12 Gaucher disease experts. CGI-S scores of 1–1.5 indicate "mild," 1.5–2.0 "moderate," 2.0–2.5 "marked," and 2.5–3.0 "severe" disease. Boxes show that, for most patients at most time points, CGI-S scores in each half-point range correspond with a 3-point range of GD1-DS3 scores.

Interrater and intrarater reliability

The intraclass correlation coefficient between any two physicians in scoring 10 patients using the GD1-DS3 was 0.97 (Cohen's kappa), indicating no significant difference between raters and a high level of concordance. Assessment of intrarater reliability, although initially planned, was not carried out.

Correlation with existing scales (construct and criterion validity)

Figure 2 represents the Pearson correlation of the GD1-DS3 with CGI-S scores from the validation conferences. In general, patients with "mild" disease as assessed by the CGI-S had GD1-DS3 scores <3, "moderate" disease correlated with DS3 scores of 3 to 6, "marked" disease 6 to 9, and "severe" disease >9. Correlation with the CGI-S was $R^2 = 0.89$ when both bone density and infiltration data were available. In the absence of these data points, the correlation was $R^2 = 0.77$.

Correlation between the CGI-S and the SSI was weak ($R^2 = 0.51$). The Pearson correlation between the SSI and the GD1-DS3 was $R^2 = 0.60$.

Minimal clinically important difference

The MCID for improvement in GD1-DS3 score was found to be a decrease of -3.17 and for worsening was +3.86. Cases scored as "no change in prognosis" by at least 75% of the physicians had changes in GD1-DS3 score that fell between these two values.

DISCUSSION

Methodological approaches for development of instruments to assess severity and monitor disease progression in any disease are necessarily complex. Rare conditions with heterogeneous clinical presentations such as GD1 pose even greater challenges. This study describes the development and validation of a new disease severity scoring system for GD1 and compares its outcomes with the only widely used GD1 severity score, the Zimran SSI,¹¹ which was developed before disease-specific treatment for GD was available. Initial estimations of validity, reliability, and theoretical feasibility suggest that it is a practicable and reliable instrument for assessing disease severity and progression.

The GD1-DS3 offers an advantage over simply monitoring individual signs and symptoms of GD1. In a disease where varying clinical manifestations may present irregularly and progress at different rates even within the same patient, a weighted global assessment is required to accurately portray individual disease status. Besides providing a conspectus of disease severity and progression, the structure of the GD1-DS3 corresponds well with published therapeutic goals for the treatment of GD1²⁴ and may facilitate monitoring of these goals as part of an integrated management approach to this chronic disease. It improves upon the Zimran SSI by incorporating modern methods for assessing bone marrow infiltration and bone mineral density.

The GD1-DS3 is also potentially useful for reliable interpatient comparisons. In diseases such as GD1 with no well-

Copyright © American College of Medical Genetics. Unauthorized reproduction of this article is prohibited

documented biomarker or singular symptom complex that adequately captures disease severity, clinical trial endpoints are often difficult to define. A validated scoring system with established reliability and sensitivity can facilitate comparison of trial groups and monitoring patient outcomes over time. It may also serve as a quantitative means to facilitate health outcomes research as well as epidemiological investigations.

Although the GD1-DS3 seems to perform well in those parameters tested, additional refinement and continued validation are required. Further evaluation of performance characteristics must be completed. For example, the comparison of the GD1-DS3 with the CGI provided one measure of criterion validity, which seems to be high. Additional evidence of criterion validity will be provided as sensitivity, specificity, and area under the curve are determined. Feasibility, although thought to be high based on the validators' assessments, also needs to be proven in the clinic.

Prospective validation exercises are typically the final test of the validity of an instrument. Such testing is currently being planned for the GD1-DS3. Predictive validity has yet to be determined. Construct validity has been partially demonstrated, as the instrument was shown to correlate very well with the CGI-S. Divergent validity, frequently ignored in validity testing, should also be established; this can be accomplished using the GD1-DS3 to assess patients with diseases with similar symptom profiles, such as rheumatoid arthritis, and comparing the scores with those obtained from patients with GD1.

This DS3 scoring system is unsuitable for use in pediatric patients with GD1. This is a significant limitation because about 26% of the International Collaborative Gaucher Group Registry population are currently younger than 18 years (Genzyme Corporation, data on file) and new patients are increasingly diagnosed in childhood. The authors believe that a separate DS3 will be necessary for pediatric patients, as disease domains that are critical in pediatric GD are not relevant to adults (e.g., those relating to growth and development). In its current form, the tool is also inapplicable to patients with neuronopathic GD. Finally, the GD1-DS3 is subject to confounding by concurrent acquired illnesses.

Although the GD1-DS3 will provide clinicians with an important tool for assessing and monitoring adult patients with GD1, its contextual interpretation is subject to the clinical judgment of the treating physician. The GD1-DS3 may provide a useful tool for evaluating a patient over time and in relation to other patients, but it is not a substitute for careful and thorough listening and observation. GD1-DS3 scores should not be expected alone to dictate treatment decisions such as initiation of treatment or dosage adjustment, particularly because therapy for GD1 is often intended to prevent development of symptoms as much as to treat existing symptoms. Clinical and laboratory assessments other than those incorporated into this instrument, as well as patient-reported outcomes, play a critical role in the effective management of GD1. Individual patients will have unique combinations of GD-related symptoms as well as comorbidities, emotional issues, and socioeconomic concerns that must be evaluated and dealt with on a case-by-case basis.

The GD1-DS3 provides a reliable method of assessing both intra- and interpatient severity of GD1 in adults. It is easy to implement and requires no assessments beyond the normal standard of care for such patients in the contemporary clinical setting. Our results demonstrate that, in its current form, the instrument is highly correlated with the clinical assessment of physicians who are expert in this field. Furthermore, we have shown that it is perceived as feasible and highly acceptable by physicians of varying specialties and nationalities who care for Gaucher patients. Definitive validation will require more extensive retrospective studies and prospective studies in different populations.

ACKNOWLEDGMENTS

The authors thank the members of the Validation Group: Hans Andersson, Roscoe Brady, Francesca Carubbi, Joel Charrow, Maja Di Rocco, Christine Eng, Paul Fernhoff, Daniel Gruskin, Paige Kaplan, Eugen Mengel, Pramod Mistry, Gregory Pastores. Katherine Lewis, MA, of Genzyme Corporation edited this manuscript.

NJW, MDC, TMC, GAG, W-LH, HM, AM, SvD, and AZ contributed to the development of the DS3 instrument, the generation of data, and the analysis of data collected from the Validation Group and the physician survey. EG and MY directed the study design and primary data analysis. CS and MY contributed to the planning of the survey, the collection of data, and the design of the instrument. EG created the initial manuscript draft and NW lead manuscript development. All authors contributed to the interpretation of results and the refinement of the manuscript and approved the final version of the manuscript.

REFERENCES

- Meikle PJ, Hopwood JJ, Clague AE, Carey WF. Prevalence of lysosomal storage disorders. JAMA 1999;281:249 –254.
- Poorthuis BJ, Wevers RA, Kleijer WJ, et al. The frequency of lysosomal storage diseases in The Netherlands. *Hum Genet* 1999;105:151–156.
- Pinto R, Caseiro C, Lemos M, et al. Prevalence of lysosomal storage diseases in Portugal. *Eur J Hum Genet* 2004;12:87–92.
- Grabowski G, Kolodny EH, Weinreb N. Gaucher disease: phenotypic and genetic variation. In: Scriver CR, Beaudet AL, Valle D, et al, editors. The online metabolic and molecular basis of inherited metabolic disease. New York: McGraw-Hill Companies, 2006. Available at: http://genetics. accessmedicine.com/mmbid/public/co_contents/toc_part16.html.
- Taddei TH, Kacena KA, Yang M, et al. The underrecognized progressive nature of N370S Gaucher disease and assessment of cancer risk in 403 patients. *Am J Hematol* 2009;84:208 –214.
- Fairley C, Zimran A, Phillips M, et al. Phenotypic heterogeneity of N370S homozygotes with type I Gaucher disease: an analysis of 798 patients from the ICGG Gaucher Registry. *J Inherit Metab Dis* 2008;31:738 –744.
- Weinreb NJ, Aggio MC, Andersson HC, et al. Gaucher disease type 1: revised recommendations on evaluations and monitoring for adult patients. *Semin Hematol* 2004;41(4 suppl 5):15–22.
- Baldellou A, Andria G, Campbell PE, et al. Paediatric non-neuronopathic Gaucher disease: recommendations for treatment and monitoring. *Eur J Pediatr* 2004;163:67–75.
- Charrow J, Andersson HC, Kaplan P, et al. Enzyme replacement therapy and monitoring for children with type 1 Gaucher disease: consensus recommendations. J Pediatr 2004;144:112–120.
- Weinreb NJ. Imiglucerase and its use for the treatment of Gaucher's disease. Expert Opin Pharmacother 2008;9:1987–2000.
- Zimran A, Sorge J, Gross E, Kubitz M, West C, Beutler E. Prediction of severity of Gaucher's disease by identification of mutations at DNA level. *Lancet* 1989;2:349 –352.
- Hermann G, Goldblatt J, Levy RN, Goldsmith SJ, Desnick RJ, Grabowski GA. Gaucher's disease type 1: assessment of bone involvement by CT and scintigraphy. *AJR Am J Roentgenol* 1986;147:943–948.
- Cox TM, Aerts JM, Belmatoug N, et al. Management of non-neuronopathic Gaucher disease with special reference to pregnancy, splenectomy, bisphosphonate therapy, use of biomarkers and bone disease monitoring. *J Inherit Metab Dis* 2008;31:319 –336.
- Di Rocco M, Giona F, Carubbi F, et al. A new severity score index for phenotypic classification and evaluation of responses to treatment in type I Gaucher disease. *Haematologica* 2008;93:1211–1218.
- Tugwell P, Boers M. Developing consensus on preliminary core efficacy endpoints for rheumatoid arthritis clinical trials. OMERACT Committee. *J Rheumatol* 1993;20:555–556.
- Bellamy N, Kirwan J, Boers M, et al. Recommendations for a core set of outcome measures for future phase III clinical trials in knee, hip, and hand osteoarthritis. Consensus development at OMERACT III. *J Rheumatol* 1997; 24:799–802.
- 17. Delbecq AL, Van de Ven AH, Gustafson DH. Guidelines for conducting

© 2010 Lippincott Williams & Wilkins

nominal group technique meetings. In: Delbecq AL, Van de Ven AH, Gustafson DH, editors. Group techniques for program planning: a guide to nominal group and Delphi processes. Middleton, WI: Greeenbriar Press, 1986.

- Bombardier C, Tugwell P. A methodological framework to develop and select indices for clinical trials: statistical and judgmental approaches. *J Rheumatol* 1982;9:753–757.
- Guy W. Clinical Global Impression. New clinical drug evaluation unit (ECDEU) assessment manual for psychopharmacology. Rockville, MD: National Institute of Mental Health, 1976:218 –222.
- Huber CG, Lambert M, Naber D, et al. Validation of a Clinical Global Impression Scale for Aggression (CGI-A) in a sample of 558 psychiatric patients. *Schizophr Res* 2008;100:342–348.
- Schneider LS, Olin JT. Clinical global impressions in Alzheimer's clinical trials. Int Psychogeriatr 1996;8:277–288; discussion 288–290.
- Beaton DE, Bombardier C, Katz JN, Wright JG. A taxonomy for responsiveness. J Clin Epidemiol 2001;54:1204 –1217.
- Weinreb N, Barranger J, Packman S, et al. Imiglucerase (Cerezyme) improves quality of life in patients with skeletal manifestations of Gaucher disease. *Clin Genet* 2007;71:576 –588.
- Maas M, van Kuijk C, Stoker J, et al. Quantification of bone involvement in Gaucher disease: MR imaging bone marrow burden score as an alternative to Dixon quantitative chemical shift MR imaging-initial experience. *Radiology* 2003;229:554-561.
- Pastores GM, Weinreb NJ, Aerts H, et al. Therapeutic goals in the treatment of Gaucher disease. Semin Hematol 2004;41(4 suppl 5):4 –14.