Evaluation of Reliability, Validity, and Responsiveness of the CDASI and the CAT-BM

Renato Goreshi[1,2], Joyce Okawa[1,2], Matt Rose[1,2], Rui Feng[3], Lela A. Lee[4], Christopher B. Hansen[5], Carolyn A. Bangert[6], M. Kari Connolly[7], Mark D. Davis[8], Jeff P. Callen[9], Nicole M. Fett[1,2], Steven S. Fakharzadeh[1,2], Jennie T. Clarke[10] and Victoria P. Werth[1,2]

To properly evaluate therapies for cutaneous dermatomyositis (DM), it is essential to administer an outcome instrument that is reliable, valid, and responsive to clinical change, particularly when measuring disease activity. The purpose of this study was to compare two skin severity DM outcome measures, the Cutaneous Disease and Activity Severity Index (CDASI) and the Cutaneous Assessment Tool—Binary Method (CAT-BM), with the Physician Global Assessment (PGA) as the "gold standard". Ten dermatologists evaluated 14 patients with DM using the CDASI, CAT-BM, and PGA scales. Inter- and intra-rater reliability, validity, responsiveness, and completion time were compared for each outcome instrument. Responsiveness was assessed from a different study population, where one physician evaluated 35 patients with 110 visits. The CDASI was found to have a higher inter- and intra-rater reliability. Regarding construct validity, both the CDASI and the CAT-BM were significant predictors of the PGA scales. The CDASI had the best responsiveness among the three outcome instruments examined. The CDASI had a statistically longer completion time than the CAT-BM by about 1.5 minutes. The small patient population may limit the external validity of the findings observed. The CDASI is a better clinical tool to assess skin severity in DM.

Journal of Investigative Dermatology (2012) **132,** 1117–1124; doi:10.1038/jid.2011.440; published online 5 January 2012

INTRODUCTION

Dermatomyositis (DM) is a chronic systemic autoimmune disease categorized among the idiopathic inflammatory myopathies (Dugan *et al*, 2009). DM is often associated with extramuscular and extracutaneous pathology, with involvement of the joints, heart (cardiomyopathy and conduction defects), and lungs (Iorizzo and Jorizzo, 2008). The most widely accepted classification criteria for DM has traditionally emphasized the importance of clinical, laboratory, histopathological, or electrophysiological evidence of muscle inflammation for making the diagnosis (Bohan and Peter, 1975a, b). Subtypes of DM, amyopathic and hypomyopathic DM, have been described for patients with no or minor muscle findings, respectively (Gerami *et al.*, 2006).

Characteristic inflammatory skin changes are seen in a large majority of individuals with DM (Callen and Wortmann, 2006). Nevertheless, the cutaneous manifestations of DM are among the least systemically studied aspects of the disease. This has resulted in part from the lack of validated tools to reliably determine the activity of the cutaneous manifestations of DM, especially relative to other dermatological diseases such as psoriasis and atopic dermatitis, where disease-specific skin severity outcome instruments have been used extensively (Kunz *et al*, 1997; Feldman and Kruger, 2005; Mrowietz *et al.*, 2006; Gaines and Werth, 2008). The Federal Drug Administration has developed guidelines for researchers on how to measure clinical response through measuring disease activity, disease-induced damage, the response as determined by the patient, and health-related quality of life (Guidance for Industry Systemic Lupus Erythematosus, 2010; Gaines and Werth, 2008). From these guidelines, researchers must develop an outcome instrument that will capture appropriate elements of the disease to determine clinical response.

Currently, effective treatments for the cutaneous manifestation of DM are limited. There are a number of new biological therapies that may be beneficial for patients with DM (Iorizzo and Jorizzo, 2008). There is a critical need

[1]*Philadelphia VA Medical Center, Philadelphia, Pennsylvania, USA;* [2]*University of Pennsylvania School of Medicine, Philadelphia, Pennsylvania, USA;* [3]*Center for Clinical Epidemiology and Biostatistics, University of Pennsylvania, Philadelphia, Pennsylvania, USA;* [4]*University of Colorado, Denver, Colorado, USA;* [5]*University of Utah, Salt Lake City, Utah, USA;* [6]*University of Texas, Houston, Texas, USA;* [7]*University of California, San Francisco, San Francisco, California, USA;* [8]*Mayo Clinic, Rochester, Minnesota, USA;* [9]*University of Louisville, Louisville, Kentucky, USA and* [10]*Penn State Hershey, Hershey, Pennsylvania, USA*

Correspondence: Victoria P. Werth, Department of Dermatology, Perelman Center for Advanced Medicine, Suite 1-330A, 3400 Civic Center Boulevard Philadelphia, Pennsylvania 19104, USA. E-mail: werth@mail.med.upenn.edu

Abbreviations: CAT, Cutaneous Assessment Tool; CAT-BM, Cutaneous Assessment Tool—Binary Method; CAT-MM, Cutaneous Assessment Tool—Maximum Method; CDASI, Cutaneous Dermatomyositis Disease Area and Severity Index; DM, dermatomyositis; DSSI, Dermatomyositis Skin Severity Index; ICC, intraclass correlation coefficient; PGA, Physician Global Assessment; SRM, standardized response mean; VAS, Visual Analogue Scale

Received 14 March 2011; revised 18 June 2011; accepted 11 July 2011; published online 5 January 2012

to develop optimal validated instruments to quantify organ-specific disease activity, so that the efficacy of medications can be methodically and quantitatively evaluated.

We have previously validated a cutaneous severity outcome instrument, the Cutaneous Dermatomyositis Disease Area and Severity Index (CDASI), and have shown that it may be a more effective and reliable tool compared with other outcome measures, namely the Dermatomyositis Skin Severity Index (DSSI) and the Cutaneous Assessment Tool (CAT; Klein *et al.*, 2008). To further simplify the CDASI, we have revised the original CDASI and have shown that the modified version correlates almost perfectly with the original CDASI (Yassaee *et al.*, 2010). The CAT was originally developed with similar goals to the CDASI and was found to have appropriate reliability, construct validity, and responsiveness in the juvenile DM population (Huber *et al.*, 2007, 2008a, b). Recently, the CAT has also been simplified, and has been validated in the juvenile population (Huber *et al.*, 2008a, b). The modified versions of the CAT, named CAT-Binary Method (CAT-BM) and CAT-Maximum Method (CAT-MM), stem from an alternative scoring method of the CAT. The CAT-BM has been shown to correlate almost perfectly to the original CAT (Huber *et al.*, 2008a, b). As yet, there are no studies comparing the modified CDASI and the CAT-BM for use in longitudinal clinical research.

The current study evaluates and compares the modified tools, with a goal to provide partial validation of each tool for use in the adult DM population and to determine the optimal effective research tool for measuring the severity of cutaneous disease in adult DM. The goal is to establish an appropriate tool for evaluating DM within and between studies to evaluate therapeutic responses most effectively.

RESULTS

Distribution of scores

CDASI Total and CAT-BM Total scores had a normal distribution with scores ranging from 1 to 72 and from 1 to 20, respectively (CDASI Total: mean 24.25 ± 14.67; CAT-BM Total: mean 9.24 ± 4.17).

Inter-rater reliability

Inter-rater reliability was assessed by determining the agreement between the CDASI and the CAT-BM scores from the 10 physician raters. The CDASI was found to have good inter-rater reliability among activity and total scores and moderate inter-rater reliability in damage scores, meaning the scores among physicians were in good accordance with one another among activity and total scores and in moderate accordance with one another among damage scores. Contrastingly, the CAT-BM was found to have moderate inter-rater reliability in activity scores and poor inter-rater reliability among damage and total scores. The CDASI had the best inter-rater reliability overall when compared with the CAT-BM and PGA scales (Activity: CDASI 0.748, CAT-BM 0.516, PGA Activity 0.721, PGA Activity Likert 0.653; Damage: CDASI 0.563, CAT-BM 0.340, PGA Damage 0.506, PGA Damage Likert 0.542; Total CDASI 0.726, CAT-BM 0.432, PGA Overall 0.632, PGA Overall Likert 0.694; Table 1).

Table 1. Assessment of inter-rater reliability

Inter-rater reliability	ICC	95% CI
CDASI		
Activity	0.748	0.553–0.895
Damage	0.563	0.358–0.785
Total	0.726	0.527–0.883
CAT-BM		
Activity	0.516	0.318–0.751
Damage	0.34	0.172–0.602
Total	0.432	0.241–0.687
PGA—Activity	0.721	0.540–0.877
PGA—Activity Likert	0.653	0.446–0.860
PGA—Damage	0.506	0.313–0.743
PGA—Damage Likert	0.542	0.329–0.797
PGA—Overall	0.632	0.422–0.835
PGA—Overall Likert	0.694	0.486–0.889

Abbreviations: CAT-BM, Cutaneous Assessment Tool—Binary Method; CDASI, Cutaneous Dermatomyositis Disease Area and Severity Index; CI, confidence interval; ICC, intraclass correlation coefficient; PGA, Physician Global Assessment.

Intra-rater reliability

Intra-rater reliability measures the degree of agreement of multiple outcome scores performed by a single physician. It was assessed by determining the agreement between initial and repeat scores, using the intraclass correlation coefficient (ICC), for each outcome instrument, as well as determining the significance of a difference between mean initial scores and mean repeat scores for each outcome instrument. The CDASI was found to have an almost perfect intra-rater reliability between activity and total scores and good intra-rater reliability with damage scores (ICC: Activity 0.868; Damage 0.800; Total 0.903). No significant difference between mean initial and mean repeat activity, damage, and total scores was found (mean difference: Activity 0.00, $P=1.00$; Damage 0.40, $P=0.728$; Total −0.40, $P=0.541$). The CAT-BM was found to have good intra-rater reliability between activity, damage scores, and total scores (ICC: Activity 0.714; Damage 0.792; Total 0.800). No significant difference between mean initial and mean repeat activity, damage, and total scores was found (mean difference: Activity 0.2, $P=0.713$; Damage 0.35, $P=0.496$; Total −0.15, $P=0.634$). PGA scales were found to have almost perfect intra-rater reliability in all assessments except for PGA Activity Likert and PGA Damage Likert (ICC: 0.737 and 0.708, respectively). There was also a significant difference between initial and repeat mean scores for PGA Overall and

Table 2. Intra-rater reliability—ICC and mean differences of initial and repeat scores

	ICC	95% CI	Mean difference between initial and repeat score	*P*-value
CDASI				
Activity	0.868	0.696–0.946	0	1.000
Damage	0.8	0.564–0.916	0.40	0.728
Total	0.903	0.770–0.960	−0.40	0.541
CAT-BM				
Activity	0.714	0.409–0.876	0.20	0.713
Damage	0.792	0.547–0.912	0.35	0.496
Total	0.8	0.561–0.916	−0.15	0.634
PGA—Activity	0.911	0.788–0.964	0.04	0.859
PGA—Activity Likert	0.737	0.399–0.892	−0.24	0.021
PGA—Damage	0.814	0.587–0.922	0.09	0.815
PGA—Damage Likert	0.708	0.382–0.877	−0.05	0.716
PGA—Overall	0.887	0.673–0.958	0.63	0.019
PGA—Overall Likert	0.875	0.703–0.950	0	1.000

Abbreviations: CAT-BM, Cutaneous Assessment Tool—Binary Method; CDASI, Cutaneous Dermatomyositis Disease Area and Severity Index; CI, confidence interval; ICC, intraclass correlation coefficient; PGA, Physician Global Assessment.

PGA Activity Likert (mean difference: PGA Overall 0.63, $P=0.019$; PGA Activity Likert −0.24, $P=0.021$; Table 2).

Construct validity

Validity was assessed for the CDASI and the CAT-BM using a linear mixed model. Both the CDASI and the CAT-BM were found to be significant predictors of the compared ''gold standard'', the PGA scales using both the Visual Analogue Scale (VAS) and the Likert scale (all $P\leqslant 0.001$ among total, activity, and damage scores; Table 3), indicating that both the CDASI and the CAT-BM were good predictors of both the VAS and the Likert PGA scales.

As another means to assess construct validity and linearity, the CDASI and CAT-BM scores were grouped by Likert scores. All CDASI and CAT-BM mean scores (Total, Activity, and Damage) expressed statistically significant distinct values when grouped by Likert scores (all *P*-values $\leqslant 0.001$; Table 4), reaffirming that both tools are good predictors of the Likert PGA scales. Furthermore, both the CDASI and CAT-BM expressed a significant, near-perfect fit for linearity with all coefficient of determination values, or r^2 values $\geqslant 0.947$ (highest $P=0.026$).

Content validity

All the physicians felt that the CDASI was complete, although one physician noted that it may be useful to have a mechanism to capture lipoatrophy from panniculitis in patients. Nine of the ten physicians felt that the CAT-BM was complete. One physician felt that the CAT-BM did not adequately assess the scalp.

Responsiveness

Responsiveness was measured using the standardized response mean (SRM), defined as the ratio of the mean of the differences (i.e., CDASI and CAT-BM scores before and after a clinical change were noted) between two time points to the standard deviation of the differences. The CDASI had the highest SRM among outcome instruments (SRM: CDASI 1.25; CAT-BM 0.93; PGA Activity 1.03; PGA Activity Likert 0.61). The CDASI was the only instrument to have an SRM >1, indicating that the mean change between visits was greater than the standard deviation change between visits. As mentioned above, the CDASI had the highest intra-rater reliability among all compared outcome instruments (Table 2).

Completion time

The CDASI had a statistically longer completion time than the CAT-BM (completion time: CDASI 4.76 minutes; CAT-BM 3.19 minutes; $P<0.001$) with a mean time difference of 1.58 minutes (95% confidence interval 1.18–1.97 minutes).

Physician exit questionnaire

Six of the ten physicians felt that the CDASI would be more easily incorporated in a clinical setting than the CAT-BM. Those who preferred the CDASI mentioned the likelihood that it would be a more effective instrument to assess responsiveness, as well as the order in which the anatomical locations were organized. Contrastingly, those who preferred the CAT-BM stated that it was a quicker instrument to complete. Six of the ten physicians felt that the CAT-BM was less difficult to use. Those who preferred the CAT-BM mentioned that it was quicker to complete, whereas those who preferred the CDASI stated that the CAT-BM was ''poorly organized'' and that they would need to ''jump around'' while completing it. All the 10 physicians felt that the CDASI was a better instrument to grade skin severity and improvement over time. Physicians commented that the CDASI measures the ''degree of intensity of an eruption,'' whereas a ''binary [method] would not be helpful in estimating response to treatment'' and would ''need to have complete resolution to capture change.'' Furthermore, one physician commented that the CAT-BM included livedo reticularis in its scoring, which ''would not be expected to improve with most therapy.''

DISCUSSION

Validated outcome measures have an important role in standardizing patient care and in developing reliable clinical trials by objectively measuring the severity of disease. The scientific method states the importance of attaining reproducible results. An outcome measure, therefore, must also be reproducible in order to adequately function in future clinical trials. The importance of an outcome measure's reliability,

Table 3. Assessment of construct validity between the CDASI and the CAT-BM

	CDASI Activity	CAT-BM Activity		CDASI Activity	CAT-BM Activity
PGA Activity			**PGA Activity Likert**		
Parameter estimate	1.96	0.61	**Parameter estimate**	3.8	1.52
SE	0.299	0.09	**SE**	1.07	0.34
F	42.97	48.12	**F**	12.56	19.66
P	<0.001	<0.001	***P***	0.001	<0.001
	CDASI Damage	**CAT-BM Damage**		**CDASI Damage**	**CAT-BM Damage**
PGA Damage			**PGA Damage Likert**		
Parameter estimate	0.72	0.51	**Parameter estimate**	1.74	1.6
SE	0.1	0.06	**SE**	0.33	0.18
F	49	80.68	**F**	28.3	78.12
P	<0.001	<0.001	***P***	<0.001	<0.001
	CDASI Total	**CAT-BM Total**		**CDASI Total**	**CAT-BM Total**
PGA Overall			**PGA Overall Likert**		
Parameter estimate	2.14	0.96	**Parameter estimate**	6.65	2.96
SE	0.36	0.12	**SE**	1.31	0.45
F	35.65	60.93	**F**	25.68	42.4
P	<0.001	<0.001	***P***	<0.001	<0.001

Abbreviations: CAT-BM, Cutaneous Assessment Tool—Binary Method; CDASI, Cutaneous Dermatomyositis Disease Area and Severity Index; PGA, Physician Global Assessment.
Data derived by using PGA values as a fixed-effect covariate and physician subject # and patient subject # as a random-effect factor. *P*-values <0.05 indicate that the outcome instrument score is a significant predictor of the ''gold standard'', its PGA counterpart. Comparisons between the CDASI and the CAT-BM cannot be made.

which measures reproducibility, is clearly important and is necessary for attaining validity (Klein *et al*., 2008; Downing, 2004). ICC values were compared via the method described by Steel *et al*. (1997). Although post-hoc power analysis showed that the difference in ICC scores did not reach statistical significance, there is a trend that the CDASI has good inter-rater reliability with regard to its Activity and Total measurements, whereas the CAT-BM has moderate and poor inter-rater reliability for its Activity and Total measurements, respectively (Table 1). Likely, the nature of the instruments lends the CDASI to having a higher inter-rater reliability, even though the CAT-BM is a binary instrument. For example, an item on the CAT-BM that was seen to have a large standard deviation among raters was the item scoring the presence of non-sun exposed erythema. As the CDASI has five to six items that would qualify as non-sun exposed erythema in addition to a larger number of items contributing to the activity score, it lends itself to having an intrinsically high inter-rater reliability, as one disagreement among physicians would have less of an impact on the overall reliability than in the CAT-BM. In addition, it is also possible that as the CDASI specifically goes through all anatomical parts, it gives more ''pressure'' to the rater to look through all the parts more efficiently than in the CAT-BM. Third, the ambiguousness of certain question items in the CAT-BM may have contributed to a lower reliability. For example, the items scoring the presence of cuticular overgrowth or subcutaneous edema were seen to have a large standard deviation among raters. Although the CDASI may not be a binary system, the measures of activity that it scores (erythema, scale, and erosions) are defined more clearly among physicians than certain measures of activity in the CAT-BM. Notably, the inter-rater reliability among activity scores in the initial study exploring the CAT-BM (Huber *et al*. 2008a, b) reports an ICC score of 0.6 (95% confidence interval 0.06–0.83), contrasting to our reported value of 0.34. Although our value of 0.34 lies within the 95% confidence interval making statistical variability the most likely cause for the difference, the differing patient populations between the studies (adult vs. juvenile) may have also had a role.

Interestingly, inter-rater reliability of damage measurements was lower in both the CDASI, the CAT-BM, and PGA scales (Table 1, ICC: CDASI Damage 0.563; CAT-BM 0.340; PGA Damage 0.506, PGA Damage Likert 0.542). This is consistent for other outcome instruments that contain a damage subscore such as the CAT and the previous version of the CDASI, suggesting that physicians have difficulty agreeing with one another in their assessment of damage (Shrout and Fleiss, 1979). It was noted that in the physician training session, the concept of poikiloderma varied among physicians. In addition, in a previous study, agreement of a physician's perception of poikiloderma was poor as well (Klein *et al*., 2008). Poikiloderma accounts for almost half, less than 10%, and theoretically 100% of the maximum damage score in the CDASI, the CAT-BM, and the PGA Damage scales, respectively. This suggests that there is

Table 4. Determination of differences among CDASI and CAT-BM scores when grouped by Likert score with linear trend of means

	CDASI Activity						CAT-BM Activity					
Likert score	*N*	Mean	SD	*P*	r^2	Linear trend *P*	*N*	Mean	SD	*P*	r^2	Linear trend *P*
0	3	0.33	0.58	<0.001	0.985	0.001	3	0.67	0.58	<0.001	0.947	0.005
1	79	11.44	6.96				79	5.04	2.36			
2	40	25.65	8.95				40	8.00	1.96			
3	13	39.38	7.25				13	9.77	1.79			
4	2	42.50	0.71				2	10.50	0.71			
Total	137	18.45	12.52				137	6.34	2.91			
	CDASI Damage						CAT-BM Damage					
0	16	0.50	0.89	<0.001	0.949	0.005	16	0.19	0.54	<0.001	0.964	0.003
1	69	4.20	2.93				69	2.33	1.60			
2	39	8.51	2.85				39	4.15	1.95			
3	13	10.62	3.38				13	5.69	1.65			
Total	137	5.61	4.07				137	2.92	2.20			
	CDASI Total						CAT-BM Total					
1	71	14.21	7.38	<0.001	0.950	0.026	71	6.85	3.11	<0.001	0.993	0.004
2	47	30.26	10.00				47	10.94	3.26			
3	16	49.50	9.48				16	14.63	2.92			
4	1	51.00	—				1	16.00	—			
Total	135	24.25	14.84				135	9.26	4.18			

Abbreviations: CAT-BM, Cutaneous Assessment Tool—Binary Method; CDASI, Cutaneous Dermatomyositis Disease Area and Severity Index.

another factor, perhaps an inherent limitation of the outcome measure, explaining the poor, and lower, inter-rater reliability of the CAT-BM when compared with the CDASI or PGA Damage scales.

The intra-rater reliability of the CDASI was almost perfect in activity and total scores and good across damage scores. The CAT-BM had a lower intra-reliability across activity, damage, and total scores with good intra-rater reliability in all realms (Table 2). Although this shows a trend that the CDASI has a better intra-rater reliability, post-hoc power analysis showed that the difference did not reach statistical significance.

Although an outcome instrument may be reliable, if it does not have adequate construct validity, or the ability to measure what it has been designed to measure effectively, then its usefulness is limited. Both the CDASI and the CAT-BM were shown to be significant predictors of PGA scales, which is the ‘‘gold standard’’, and thus to have good construct validity. Although both the CDASI and the CAT-BM were found to have good content validity as stated above, a physician noted that the CAT-BM did not sufficiently assess scalp disease, which can be very troublesome for patients and was found in over 80% in the DM population (Kasteler and Callen, 1994; Tilstra *et al.*, 2009).

It is also important for an outcome instrument to be able to capture the disease state of patients at the extremes of disease. This is particularly important in patients with extreme disease activity. In this study, the maximum CDASI Activity and CAT-BM Activity score reached was 61 (61% of maximum activity score) and 14 (82% of maximum activity score), respectively. This suggests that the CAT-BM may be more prone to reach its maximum limit faster than the CDASI, and therefore not be able to capture differences in disease activity in more severe patients.

To implement an outcome instrument for the use of clinical trials, it is essential that it be able to measure changes in disease severity. The CDASI had the best responsiveness when compared with CAT-BM and PGA scales. Furthermore, all physicians anticipated that the CDASI would be a more effective response tool than the CAT-BM. This was not a surprising result, as shown by many of the physician rater comments, predicting that the CAT-BM would have this limitation as it only documents presence or absence of a certain measure, whereas the CDASI documents the degree of severity of a certain measure.

Another important factor when comparing outcome instruments is its completion time. Even a tool that is reliable and valid but takes too long to complete would not be practical in a clinical research setting. Although the CAT-BM took significantly less time to complete than the CDASI (Mean completion time: CAT-BM 3.19 minutes; CDASI 4.76 minutes; $P<0.001$), the mean difference in completion time was about 90 seconds and may not be practically relevant.

There were limitations to the study. First, as the patient population was relatively small, the external validity of our findings may be limited. Many results were not able to reach statistical significance because of a small patient population. For example, post-hoc power analysis showed that a patient population of 60 would be needed to reach statistical significance between the ICC scores. Second, the relatively small patient population may have allowed the physician raters to recall how they evaluated a patient when completing their repeat evaluation. This could potentially raise the intra-rater reliability from its true value. To minimize this impact, physicians were asked to perform their repeat evaluation on a patient they had evaluated during the morning session, thus minimizing a likelihood of recall. Third, as the study session lasted about 7 hours, it is possible that the physicians may have experienced fatigue that may have impacted their patient evaluation. This was minimized by offering snacks and lunch during the day and allowing physicians to rate patients at their own pace. Fourth, five of the ten physician participants have used both the CAT and the original version of the CDASI previously, which may have falsely elevated the reliability and validity scores in both instruments, as many physicians had increased familiarity with both the instruments. Regardless of the limitations above, we can conclude that the CDASI appears to be a more effective tool than the CAT-BM in evaluating cutaneous severity in DM.

MATERIALS AND METHODS

This study has been approved by the local Institutional Review Board. The Declaration of Helsinki Principles protocols was adhered and physician and patient participants gave their written, informed consent before study initiation.

Physician participants

A total of 10 dermatology-boarded physicians were invited to participate in the 1-day study at the Hospital of the University of Pennsylvania. Physicians were given the CDASI and the CAT-BM, as well as corresponding literature before the study session day so that they may better familiarize themselves with the tools. On the study session day, before initiating the study, the physicians were given a training session with visual examples in order to score all study instruments correctly. Adequate time was given to the physicians to address any questions and/or clarifications they may have had regarding the outcome instruments.

Patient participants

A total of 14 patients with the clinical and/or pathological evidence of DM were invited to participate in the study at the Hospital of the University of Pennsylvania. The patients represented a wide spectrum of diseases. The patient population consisted of 14 Caucasians, 3 men, 11 women, with varying degrees of muscle and cutaneous involvement (noted to have PGA Activity scores ranging from 0 to 9.3 with a mean of 3.2 ± 2.8; PGA Damage scores ranging from 0 to 9.4 with a mean of 2.8 ± 2.6; and PGA Overall scores ranging from 0.2 to 9.2 with a mean of 3.4 ± 2.5). Average age of participants was 53 ± 16. Average duration of disease among patients was not recorded.

Study design

The study day was divided into Session 1 and Session 2. Each physician was given a randomized number from 1 to 10 and consequently a folder corresponding to their number. On the basis of the assigned number, physicians were divided into two groups of five physicians—Group 1Ph and Group 2Ph. One physician group contained folders with packets of each outcome instrument in the order of CDASI, CAT-BM, and PGA scales for Session 1 and packets of each outcome instrument in the order of CAT-BM, CDASI, and PGA scales for Session 2. The remaining physician group contained folders with a reverse order of packets (i.e., CAT-BM, CDASI, and PGA scales for Session 1). All folders from both the physician groups also contained two packets of each outcome instrument for re-rates. All physicians evaluated 14 patients. All physicians also reevaluated two patients. At the end of the study session, physicians were given an exit questionnaire consisting of seven questions, each of which consisting of a short answer and four questions including a multiple-choice part. Patients were randomized and divided into two groups: Group 1P, consisting of eight patients, and Group 2P, consisting of six patients. During Session 1, Group 1Ph evaluated Group 1P and Group 2Ph evaluated Group 2P. During Session 2, Group 1Ph evaluated Group 2P and Group 2Ph evaluated Group 1P. No more than one physician was permitted per patient encounter at any time.

CDASI

The CDASI is a one-page, partially validated outcome instrument used to determine the severity of cutaneous disease specific to DM. Total scores range from 0 to 132. Scores are divided into activity and damage, with scores ranging from 0 to 100 and 0 to 32, respectively. Neither activity nor damage is scored by percentage of body surface area involvement. Disease activity is assessed by the degree of erythema, scale, and the presence of erosions or ulcerations in 15 different anatomical locations. Disease damage is assessed by presence of poikiloderma or calcinosis in the 15 different anatomical locations. Periungual changes were scored from 0 to 2, with 0 indicating no periungual changes, 1 indicating periungual erythema, and 2 indicating visible telangectasias. Alopecia scores range from 0 to 1, with zero indicating no alopecia in the last 30 days and 1 indicating presence of alopecia in the last 30 days. Gottron's sign on the knuckles are assessed similarly to the erythema scale used in other anatomical locations. When Gottron's papules were present, the erythema score obtained on the knuckles was doubled.

CAT-BM

The CAT-BM is a one-page, normally distributed validated outcome instrument derived from an alternative scoring method of the CAT that is used to determine the severity of cutaneous disease in DM. Total scores range from 0 to 28 and 0 to 17 for activity and from 0 to 11 for damage. Neither activity nor damage is scored by percentage of body surface area involvement. Activity scores are based on the presence of erythema in seven different anatomic areas and presence of other characteristic DM lesions. Secondary changes such as scales, erosions, or necrosis are not captured. Disease damage is scored by the presence of atrophy or dyspigmentation without erythema in the same seven different anatomic areas, as well as presence of poikiloderma, calcinosis, lipoatrophy, or a depressed scar anywhere on the body.

Assessment of inter- and intra-rater reliability

To assess intra-rater reliability, after a physician participant had completed all patient encounters, they were asked to reevaluate two patients whom they had seen during the morning session (to minimize physician recollection of scoring). Although physicians arbitrarily decided which patient to re-rate on the basis of patient availability, it was ensured that no patient would be re-rated more than twice. Inter-rater reliability was used to assess accordance of scores among physicians. All physicians re-rated two patients. Inter-rater reliability was determined by the 10 physicians who evaluated all the 14 patients. Physicians also recorded the time to complete each instrument for each patient encounter.

Validation measures

To assess and compare validity among different outcome instruments, three validation measures were used: (1) the Overall Skin-Physician Global Assessment (PGA Overall), (2) the Skin Activity-Physician Global Assessment (PGA Activity), and (3) the Skin Damage-Physician Global Assessment (PGA Damage). Scores were captured using VAS and Likert scales. The VAS is a continuous scale ranging from 0 to 10, where 10 represents extremely active disease. The Likert scale ranges from 0 to 4, where 4 represents extremely severe disease.

Assessment of validity

Specifically, convergent construct validity was determined by comparing the Skin Activity-PGA to the activity scores of the activity subscore of the outcome instruments, comparing the Skin Damage-PGA to the damage subscore of the outcome instruments, and comparing the Overall Skin-PGA to the overall score of the outcome instruments. Convergent construct validity refers to the degree one measure (i.e., the CDASI or the CAT-BM) correlates to another measure (i.e., the corresponding PGA) that it theoretically should correlate with. The PGAs were also used to determine whether either of the outcome instruments was skewed to any direction, which could potentially limit the usefulness in longitudinal studies. Content validity was determined by administrating the Physician Exit Questionnaire, which includes the question ''Was there any information missing from any of the measures that you feel should be added?''

Responsiveness

Responsiveness was assessed from prospective visit data collected separately from the inter- and intra-rater validation studies. This included assessments of the CDASI, CAT-BM, and PGA scale scores, as well as an overall evaluation from the physician as to whether the patient had improved, worsened, or had no change from their previous research visit. A total of 35 patients with a cumulative 110 visits were obtained from this data source. There were 27 visits in which a clinical change was noted. The largest clinical change per patient, defined as the largest difference in the PGA-Activity score between two consecutive visits, was included in the analysis. The SRM was used to determine responsiveness for the CDASI and the CAT-BM. The SRM measures the ratio of the mean of the differences (i.e., CDASI and CAT-BM scores before and after a clinical change was noted) between two time points to the standard deviation of the differences. The absolute mean change was used between visits to account for improvement and worsening of disease. This approach has been used in the past (Beaton *et al.*, 1997; Ruperto *et al.*, 2010).

Statistical methods

Statistical analyses were performed using statistical programs STATA and SPSS. Inter-rater reliability was determined by ICC, type ICC (2,1), via Shrout and Fleiss convention (Shrout and Fleiss, 1979). Previous research has dictated that an ICC between 0.5 and 0.7 to be moderate, between 0.70 and 0.81 to be good, and an ICC $\geqslant$0.81 to be almost perfect (Landis and Koch, 1977; Klein *et al.*, 2008). Intra-rater reliability was determined by ICC (2,1) and paired, two-tailed *t*-test comparing mean scores between initial and repeat scores of each instrument. Construct validity was assessed by testing the association between outcome measure (CDASI or CAT-BM) and the corresponding validation measure. Because each patient and each physician had repeated measures, we used a linear mixed model for this test, adjusting for within-patient and within-physician variations. Other covariates, such as age and gender, were not seen to have an influence. Physician subject # and patient subject # were placed as random-effect factors, whereas PGA scores were placed as a fixed-effect covariate. Likert scores were also used as an additional means to assess construct validity. Differences in CDASI and CAT-BM scores when grouped by corresponding Likert scores were evaluated using one-way analysis of variance. Linear regression was also used on mean CDASI and CAT-BM scores of each Likert group to determine linearity.

CONFLICT OF INTEREST

The authors state no conflict of interest.

ACKNOWLEDGMENTS

This material is based on work supported by Celgene Corporation, the Department of Veterans Affairs Veterans Health Administration, Office of Research and Development, Biomedical Laboratory Research and Development, and by the National Institutes of Health (NIH K24-AR 02207) (VPW).

REFERENCES

Beaton DE, Hogg-Johnson S, Bombardier C (1997) Evaluating changes in health status: reliability and responsiveness of five generic health status measures in workers with musculoskeletal disorders. *J Clin Epidemiol* 50:79–93

Bohan A, Peter JB (1975a) Polymyositis and dermatomyositis (first of two parts). *N Engl J Med* 292:344–7

Bohan A, Peter JB (1975b) Polymyositis and dermatomyositis (second of two parts). *N Engl J Med* 292:403–7

Callen JP, Wortmann RL (2006) Dermatomyositis. *Clin Dermatol* 24:363–73

Guidance for Industry Systemic Lupus Erythematosus—Developing Medical Products for Treatment. US Department of Health and Human Services, Food and Drug Administration, Center for Drug Evaluation and Research, Center for Biologics Evaluation and Research, Center of Radiological Health. June 2010. http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM072063.pdf

Downing SM (2004) Reliability: on the reproducibility of assessment data. *Med Educ* 38:1006–12

Dugan EM, Huber AM, Miller FW *et al.* (2009) Review of the classification and assessment of the cutaneous manifestations of idiopathic inflammatory myopathies. *Dermatol Online J* 15:2

Feldman SR, Krueger GG (2005) Psoriasis assessment tools in clinical trials. *Ann Rheum Dis* 64(Suppl 2):65–8; discussion 69–73

Gaines E, Werth VP (2008) Development of outcome measures for autoimmune dermatoses. *Arch Dermatol Res* 300:3–9

Gerami P, Schope JM, McDonald L *et al.* (2006) A systematic review of adult-onset clinically amyopathic dermatomyositis (dermatomyositis sine' myositis): a missing link within the spectrum of the idiopathic inflammatory myopathies. *J Am Acad Dermatol* 54:597–613

Huber AM, Dugan EM, Lachenbruch PA *et al.* (2007) The cutaneous assessment tool: development and reliability in juvenile idiopathic inflammatory myopathy. *Rheumatology (Oxford)* 46: 1606–11

Huber AM, Dugan EM, Lachenbruch PA *et al.* (2008a) Preliminary validation and clinical meaning of the cutaneous assessment tool in juvenile dermatomyositis. *Arthritis Rheum* 59:214–21

Huber AM, Lachenbruch PA, Dugan EM *et al.* (2008b) Alternative scoring of the cutaneous assessment tool in juvenile dermatomyositis: Results using abbreviated formats. *Arthritis Rheum* 59:352–6

Iorizzo III LJ, Jorizzo JL (2008) The treatment and prognosis of dermatomyositis: an updated review. *J Am Acad Dermatol* 59:99–112

Kasteler JS, Callen JP (1994) Scalp involvement in dermatomyositis. Often overlooked or misdiagnosed. *JAMA* 272:1939–41

Klein RQ, Bangert CA, Costner M *et al.* (2008) Comparison of the reliability and validity of outcome instruments for cutaneous dermatomyositis. *Br J Dermatol* 159:887–94

Kunz B, Oranje AP, Labrèze L *et al.* (1997) Clinical validation and guidelines for the SCORAD index: consensus report of the European Task Force on Atopic Dermatitis. *Dermatology* 195:10–9

Landis JR, Koch GG (1977) The measurement of observer agreement for categorical data. *Biometrics* 33:159–74

Mrowietz U, Elder JT, Barker J (2006) The importance of disease associations and concomitant therapy for the long-term management of psoriasis patients. *Arch Dermatol Res* 298:309–19

Ruperto N, Bazso A, Pistorio A *et al.* (2010) Agreement between multidimensional and renal-specific response criteria in patients with juvenile systemic lupus erythematosus and renal disease. *Clin Exp Rheumatol* 28:424–33

Shrout PE, Fleiss JL (1979) Intraclass correlations: uses in assessing rater reliability. *Psychol Bull* 86:420–8

Steel RD, Torrie JH, Dickey DA (1997) *Principles and Procedures of Statistics: A Biometrical Approach*. McGraw Hill Book, New York, 297–9

Tilstra JS, Prevost N, Khera P *et al.* (2009) Scalp dermatomyositis revisited. *Arch Dermatol* 145:1062–3

Yassaee M, Fiorentino D, Taylor L *et al.* (2010) Modification of the cutaneous dermatomyositis disease area and severity index, an outcome measure instrument. *Br J Dermatol* 162:669–73