Published in final edited form as:
Biol Blood Marrow Transplant. 2011 November ; 17(11): 1619–1629. doi:10.1016/j.bbmt.2011.04.002.

A Multi-Center Pilot Evaluation of the National Institutes of Health Chronic Graft-versus-Host Disease (cGVHD) Therapeutic Response Measures: Feasibility, Inter-rater Reliability, and Minimum Detectable Change

Sandra A. Mitchell[1], **David Jacobsohn**[2], **Kimberly E. Thormann Powers**[3], **Paul A. Carpenter**[4], **Mary E.D. Flowers**[4], **Edward W. Cowen**[5], **Mark Schubert**[4], **Maria Turner**[5], **Stephanie J. Lee**[4], **Paul Martin**[4], **Michael R. Bishop**[5], **Kristin Baird**[5], **Javier Bolaños-Meade**[7], **Kevin Boyd**[3], **Jane M. Fall-Dickson**[8], **Lynn H. Gerber**[6], **Jean-Pierre Guadagnini**[9], **Matin Imanguli**[5], **Michael C. Krumlauf**[6], **Leslie Lawley**[3], **Li Li**[6], **Bryce B. Reeve**[10], **Janine Austin Clayton**[11], **Georgia B. Vogelsang**[7], and **Steven Z. Pavletic**[5]

[1] Division of Cancer Control and Population Sciences, National Cancer Institute, National Institutes of Health, Bethesda, MD

[2] Division of Blood and Marrow Transplantation, Center for Cancer and Blood Disorders, Children's National Medical Center, Washington, DC

[3] Cancer and Blood Diseases, Children's Memorial Hospital, Chicago, IL

[4] Division of Clinical Research, Fred Hutchinson Cancer Research Center, and the University of Washington School of Medicine, Seattle, WA

[5] Center for Cancer Research, National Cancer Institute, National Institutes of Health, Bethesda, MD

[6] Clinical Center, National Institutes of Health, Bethesda, MD

[7] Bone Marrow Transplantation, Sidney Kimmel Comprehensive Cancer Center, Johns Hopkins University, Baltimore, MD

[8] Symptom Management Branch, National Institute of Nursing Research, Bethesda, MD

[9] National Institute of Dental and Craniofacial Research, National Institutes of Health, Bethesda, MD

[10] Lineberger Cancer Center and the Department of Health Policy and Management, University of North Carolina at Chapel Hill, Chapel Hill, NC

[11] National Eye Institute, National Institutes of Health, Bethesda, MD

Abstract

The lack of standardized criteria for measuring therapeutic response is a major obstacle to the development of new therapeutic agents for chronic graft-versus-host disease (cGVHD). National

Address correspondence to: Sandra A. Mitchell, PhD, CRNP, Outcomes Research Branch, Applied Research Program, Division of Cancer Control and Population Sciences, National Cancer Institute, 6130 Executive Blvd, MSC 7344, EPN 4020, Bethesda, MD, 20892, 301-435-6750, mitchlls@mail.nih.gov.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Institutes of Health (NIH) consensus criteria for evaluating therapeutic response were published in 2006. We report the results of four consecutive pilot trials evaluating the feasibility and estimating the inter-rater reliability and minimum detectable change of these response criteria.

Hematology-oncology clinicians with limited experience in applying the NIH cGVHD response criteria (n=34), participated in a 2.5 hour training session on response evaluation in cGVHD. Feasibility and inter-rater reliability between subspecialty cGVHD experts and this panel of clinician raters were examined in a sample of 25 children and adults with cGVHD. The minimum detectable change was calculated using the standard error of measurement.

Clinicians' impressions of the brief training session, the photo atlas, and the response criteria documentation tools were generally favorable. Performing and documenting the full set of response evaluations required a median of 21 minutes (range 12 to 60 minutes) per rater. The Schirmer tear test required the greatest time of any single test (median 9 minutes). Overall, inter-rater agreement for skin and oral manifestations was modest, however, in the third and fourth trials, the agreement between clinicians and experts for all dimensions except movable sclerosis approached satisfactory values. In the final two trials, the threshold for defining change exceeding measurement error was 19–22% body surface area (BSA) for erythema, 18–26% BSA for movable sclerosis, 17–21% BSA for nonmovable sclerosis, and 2.1–2.6 points on the 15 point NIH Oral cGHVD scale. Agreement between clinician-expert pairs was moderate to substantial for the measures of functional capacity and for the gastrointestinal and global cGVHD rating scales.

These results suggest that the NIH response criteria are feasible for use, and these reliability estimates are encouraging, because they were observed following a single 2.5 hour training session given at multiple transplant centers, with no opportunity for iterative training and calibration. Research is needed to evaluate inter- and intra-rater reliability in larger samples, and to evaluate these response criteria as predictors of outcomes in clinical trials.

Keywords

Chronic graft-versus-host disease; response criteria; inter-rater reliability; minimum detectable change

Introduction

Although new treatment approaches for chronic graft-versus-host disease (cGVHD) are emerging, progress in the development of new therapies has been limited by the absence of criteria for evaluating responses that are reliable, valid, sensitive to change, and accepted for use in clinical trials[1, 2]. In 2006, the National Institutes of Health (NIH) Consensus Development Project on Criteria for Clinical Trials in Chronic GVHD proposed a series of measures for evaluating therapeutic response in cGVHD[3]. The NIH cGVHD therapeutic response criteria are comprised of core and ancillary measures, and include both clinician-assessed and patient-reported components. The core measures include organ-specific measures of skin, mouth, and eye involvement, as well as global ratings of cGVHD severity from both clinician and patient perspectives[3]. The patient-reported outcome of cGVHD symptom bother is included as a core measure in the response criteria, and is evaluated using the Lee cGVHD Symptom Scale[4]. Ancillary measures of response include performance-based measures of physical function such as grip strength and 2-minute walk time, and a variety of patient reported-outcomes including functional status and health-related quality of life. However, the feasibility and reliability of these consensus criteria for use in evaluating adults and children with a range of cGVHD manifestations have not been defined. Evaluation of the concordance between subspecialty cGVHD experts and transplant clinicians in scoring measures of response represented the next logical step in exploring the reliability and validity of the NIH response criteria. Determining if clinicians' scoring of response reproduces the scoring of an expert is fundamental to further development of the criteria and their application in prospective clinical trials[5].

We conducted four pilot studies to (i) explore the feasibility of using the NIH cGVHD response criteria for evaluating adult and pediatric patients with cGVHD, and (ii) develop preliminary estimates of the inter-rater reliability and minimum metrically detectable difference between subspecialty experts and a group of hematology-oncology clinicians with limited experience in cGVHD who had received a single educational session about the NIH response criteria. A secondary objective was to develop and evaluate the teaching materials and data collection tools that facilitate use of the response criteria in clinical trials.

Materials and Methods

This prospective study was conducted at three sites between May 2005 and October 2006. The study was approved by the institutional review boards of the Center for Cancer Research, Bethesda, MD; Northwestern University, Chicago, IL; and Fred Hutchinson Cancer Research Center, Seattle, WA. Two of the trials were conducted in one site (trial 1 and trial 3), while trials 2 and 4 were each conducted at a single center. All participants provided written informed consent.

Hematology-oncology clinicians (attending physicians, fellows, nurse practitioners and physician's assistants) (n=34) with limited experience with cGVHD or with the NIH cGVHD response criteria participated in a 2.5 hour training session designed to provide an overview of comprehensive response evaluation in cGVHD. They also received a syllabus and a photo atlas illustrating common oral, ocular, and dermatologic manifestations of cGVHD (http://asbmt.affiniscape.com/displaycommon.cfm?an=1&subarticlenbr=29). Participants represented a spectrum of experience, ranging from hematology-oncology fellows or nurse practitioners with little experience in stem cell transplantation and cGVHD (n=17) to experienced transplant clinicians with expertise in managing patients with cGVHD (n=17). In the final two trials, all of the clinician raters had experience in stem cell transplant and cGVHD, however across all four trials, none of the clinician raters had experience applying the NIH response criteria in evaluating patients. Clinician raters were eligible to participate in only one trial. Within 24 hours following the training session, the clinician raters applied the NIH response measures in evaluating adult or pediatric patients with cGVHD. Ratings made by cGVHD experts in their respective areas (transplantation, dermatology, oral medicine, and rehabilitation) represented the standard against which inter-rater reliability was determined. The clinician and expert evaluations of each panel of patients were conducted in an ambulatory clinic setting over a period of 4–6 hours. In each of the four sequential trials, a panel of four experts was assembled to perform subspecialty-focused evaluations (dermatology, oral medicine, rehabilitation, and transplantation) of each study participant. Chronic GVHD experts were organ system/subspecialty experts with substantial experience in evaluating and managing patients with cGVHD. In each trial, these cGVHD experts evaluated all patients immediately prior to the evaluation conducted by the clinician raters (n=8, n=10, n=9 and n=7 clinician raters in trials 1 through 4, respectively). The experts were not permitted to interact with the clinician raters until all study-related procedures were concluded. To ensure that each examiner remained blinded to the ratings of other examiners, the scoring sheets were collected by the study investigators immediately after the expert or clinician raters had completed each evaluation of the patient, and the ratings of each study participant were known only to the principal investigator and the study team.

Feasibility, acceptability, and inter-rater reliability between experts in cGVHD (transplantation, dermatology, oral medicine, and rehabilitation medicine) and the 34 clinician raters were examined using 25 pediatric and adult patients with varying manifestations of cGVHD. Feasibility was evaluated by measuring the time required to perform both the total evaluation and specific components of the evaluations. Time was considered one of the most important components of operational feasibility. A second dimension of feasibility was examined by asking the non-expert clinicians to rate the extent to which (i) the training session provided the skills necessary to conduct the response evaluations, and (ii) the measures and data capture forms were acceptable and easy to use. Reliability estimates from each trial and feedback from participants were used to further modify the therapeutic response measures and the training tools, and these revised materials were then tested in subsequent trials.

Statistical Analysis

Descriptive statistics were used to report participants' demographic and clinical characteristics, and the feasibility dimensions. Extent of agreement between clinician assessors and expert raters was quantified in several ways, as recommended by Sanchez and Binkowitz[5]. First, we created a difference score for each clinician-expert pair (difference=clinician minus expert score), and examined the distributional properties of those difference scores with Bland-Altman plots, graphing the differences against the expert score. Bland-Altman plots were used to evaluate whether the difference between the clinician and expert assessments varied as a function of the extent of cGVHD involvement. The intraclass correlation coefficient (ICC) was used to quantify concordance between each clinician-expert pair. In calculating the ICC, a single-measure, two-way mixed effects, absolute agreement model was specified where patients were interpreted as a random effect, and raters were interpreted as a fixed effect[6]. This intraclass correlation coefficient was chosen since raters were a convenience sample who, it may be argued, were more motivated and interested in cGVHD than a random selection of raters[7]. Absolute agreement measures rather than consistency agreement measures were computed since systematic differences among levels of ratings may have been relevant (that is, there may have been systematic differences when scoring lower levels of BSA involvement versus higher levels of BSA involvement)[8]. In addition, because evaluation of therapeutic response in trials typically relies on the rating provided by a single clinician, rather than combining scores among several raters, the single measures ICC, rather than the average measures ICC is reported throughout this paper[9]. According to interpretive rules for the ICC proposed by Landis and Koch, an ICC of 0.21 to 0.40 represents fair agreement, 0.41 to 0.60 represents moderate agreement, 0.61 to 0.80 substantial agreement, and 0.81 to 1.00 almost perfect agreement[10].

A further goal in the analysis was to characterize the measurement error associated with the different graders using the response criteria in evaluating patients with cGVHD. We examined this in two ways. First, we created agreement parameters [# agreements/(# agreements + # disagreements)] by reporting the percentage of clinician-expert pairs who (1) differed by less than ± 10% body surface area (BSA) for cutaneous manifestations, or (2) fell within ±1 point on the 15-point NIH Oral cGHVD scale. The criteria for these agreement parameters were extrapolated from the definitions of partial and complete response for skin and oral manifestations, as proposed by Pavletic et al.[3] and represent the minimum threshold of change that would need to be observed in order to detect partial or complete response in cutaneous or oral manifestations. We also determined the percentage of clinician-expert pairwise values that (1) fell within the 95% confidence interval of the expert's value for the functional evaluations of 2 minute walk distance and grip strength, or (2) differed by ± 1 point or less on the cGVHD global assessment scales.

Finally, to complement the interpretation of the intraclass correlation coefficients we also calculated the minimum metrically detectable change (MDC_{95}) associated with rating cutaneous and oral cGVHD manifestations. The MDC_{95} provides an anchor when interpreting change scores, because only when that change score exceeds the MDC_{95} can the researcher conclude with 95% confidence that the change represents true change and not measurement error[11]. Thus the MDC_{95} offers a data-driven approach to aid in the interpretation of change scores and sample size estimation[12]. The following formula was used to calculate the MDC_{95}: (population standard deviation × [square root (1-ICC value]) × 1.96 (standard normal score associated with a two-tailed 95% confidence interval). The population standard deviations were based on data from an ongoing cGVHD natural history study at the National Cancer Institute (N=155) (clinicaltrials.gov#NCT00331968). These means and standard deviations (erythema 8.5% ± 15.1% BSA; movable sclerosis 6.9% ± 15% BSA; nonmovable sclerosis 9.6% ± 17.8% BSA; oral 1.96 ± 2.0) are comparable to those recently reported in another series[13].

All analyses were conducted using SPSS version 17. The number of rating tasks (defined as the number of clinician-expert pair-wise comparisons) available for analysis in each trial is specified in the tables where the inter-rater reliability results are presented. Missing data are accounted for by the fact that some of the patient participants did not have sufficient endurance to tolerate complete examinations by all clinicians. When they so requested, patient participants were given short breaks from the examinations to rest, and this resulted in a small amount of missing data.

Results

Patient characteristics

The patient sample (n=14 adults; n=11 children) had a mean age of 33.5 years (Range 3 to 70 years), and was predominantly male (64%). Most had undergone stem cell transplantation for a diagnosis of acute leukemia (44%) or chronic leukemia, lymphoma or multiple myeloma (28%). Approximately two thirds had received a reduced intensity conditioning regimen (60%) and a matched sibling donor graft (60%). The stem cell source was peripheral blood in approximately 80% of the patient participants. The Karnofsky/Lansky Performance Status was determined to be greater than or equal to 80% in 72% of the patient participants.

Patients were a median of 40 months post-transplant (range 5–195 months). More than two-thirds of the sample had moderate to severe cGVHD, as defined by the NIH global severity scoring[14], with a median of 4 (range 1 to 7) organ sites involved. Based on the NIH diagnostic criteria[14], chronic GVHD onset was classified in a majority as progressive (52%), with the remainder classified as quiescent (24%) or de novo (24%). Patients were a median of 34 months (range 2–188 months) from the diagnosis of cGVHD, and a majority of the patient participants were receiving moderate (single agent/modality± prednisone ≥ 0.5mg/kg/day) (36%) or high (two or more agents/modalities ± prednisone ≥ 0.5mg/kg/day) (48%) levels of systemic immunosuppression.

The sample characteristics relative to the dimensions evaluated in the NIH cGVHD response criteria were derived from the expert raters and are presented in Table 1. As seen in the values for standard deviations about the mean, there was sizable variability observed in the scoring of involved BSA of cutaneous manifestations, including erythema, movable, and nonmovable sclerosis. Mean values for cGVHD self-reported symptom bother were moderate among participants in the first two trials, and lower levels of symptom bother were noted among participants in the third and fourth trials. In general, few participants experienced esophageal, upper, or lower GI symptoms.

Feasibility

The median total time for the novice clinician raters to complete and document all response evaluations was 21 minutes (range 12 to 60 minutes). The oral examination required a median of 3 minutes; skin evaluation a median of 6 minute;, and grip strength and walk time a median of 7 minutes. The Schirmer's tear test required the greatest amount of time of any single test (median 9 minutes). The median total time for the patients to complete all five self-report measures (MOS-Short Form 36, Human Activity Profile, Functional Assessment of Cancer Therapy, Lee Chronic GVHD Symptom Scale, and NIH Response Criteria Form B) was 15 minutes (range 8 to 22 minutes).

In terms of feasibility across all four trials, clinicians offered a generally favorable evaluation of the training session, with more than 70% reporting that the practice opportunities embedded in the training session were helpful in building the necessary skills to perform specific procedures such as grip strength evaluation and the 2-minute walk distance. At least two-thirds of clinician raters indicated that the photo atlas of cutaneous and oral cGVHD manifestations was extremely helpful in recognizing and describing cGVHD clinical features. Three quarters of participants agreed that overall the response evaluation documentation tools were clearly presented, well-organized, and easy to complete. Nonetheless, at least half of the clinician raters believed that incorporating these response assessments into their routine practice or asking community oncologists and oncology advanced practice nurses to perform these response evaluations in their clinics would be challenging.

Inter-rater reliability

Metrics of agreement between the expert and clinician raters across the four trials in evaluating cutaneous and oral cGVHD manifestations are presented in Tables 2 and 3 and in Figure 1. The median ICC was typically in the range of 0.5 to 0.6, suggesting moderate agreement, with a trend for the highest ICCs to be noted in the final two trials. The mean difference score (the absolute value of the difference between expert and clinician rater) ranged from 10.4 (S.D. ±4.2) to 21.7 (S.D. ±9.6) percent BSA for erythema, from 9.1 (S.D. ±6.5) to 17.0 (S.D. ±8.9) percent BSA for movable sclerosis, and from 7.2 (S.D. ±3.2) to 19.0 (S.D. ±6.2) percent BSA for nonmovable sclerosis. Across the four trials, the mean absolute difference score for oral manifestations ranged from 1.6 (S.D. ±0.48) to 2.7 (S.D. ±1.5).

The clinician-expert difference scores (mean difference and 95% confidence interval) for the 10 patients evaluated in the final trial plotted against the expert's scores for those patients are portrayed in Figure 1. The plots reveal the heterogeneity in the extent of cGVHD cutaneous and oral manifestations present in our sample of patients, and show that there was fair agreement between clinicians and the expert across all levels of cGVHD involvement. Relative to the expert, clinician raters tended to overestimate the BSA involved with movable sclerosis and underestimate the extent of oral involvement with cGVHD.

We also calculated agreement parameters to gauge if the proposed NIH response criteria were sensitive enough to accurately detect partial response in both cutaneous and oral evaluations (that is, capable of accurately detecting a change within 25–50% of total BSA involved with cutaneous manifestations, and for oral manifestations, within 1 point on the 15 point NIH Oral cGHVD scale)[3]. For example, in someone with BSA of 60% of skin involvement, they must demonstrate an improvement to 30% of BSA involvement if they are to be classified as partial response[3]. However, if there is 25% BSA involvement or less, only a complete resolution of all findings is considered as a response[3]. Extrapolating from this logic of accurately defining partial response, we projected that an agreement within ±

10% BSA represents a minimally acceptable limit of pair-wise agreement between clinician and expert. Following a parallel logic for the 15 point scale for rating oral manifestations, we defined acceptable agreement to have occurred when clinician and expert scores agreed within ±1 point.

As shown in Table 2, in rating erythema in the final two trials, between 54% and 61% of the time the clinician-expert pairs differed by less than ±10% body surface area. For movable sclerosis, 55% to 72% of the clinician-expert assessments agreed within ± 10% BSA; and for nonmovable sclerosis, 70% to 72% of the assessments were within ± 10% body surface of involvement. The mean absolute difference between clinician-expert pairs declined across the course of the four trials, and in the final two trials, only the parameter for erythema exceeded a mean absolute difference of 11% BSA. Between 43% and 58% of the time in the last two trials, clinicians rated oral manifestations within ± 1 point of the expert on the 15 point NIH Oral cGHVD scale (see Table 3).

Consistent with these observations, the minimum detectable change (MDC_{95}) in the last two trials ranged from 19% to 22% BSA for erythema, 18% to 26% BSA for movable sclerosis, 17% to 21% BSA for nonmovable sclerosis, and 2.1 to 2.6 points on a 15 point scale for oral cGVHD manifestations. These MDC_{95} estimates suggest that when observed change across time exceeds these estimates one can conclude with 95% confidence that the change is unlikely to have occurred due to inter-rater measurement error alone.

As summarized in Figure 2, concordance between clinician and expert pairs was quite high for the ratings of upper gastrointestinal, lower gastrointestinal and esophageal symptoms. On more than 90% of occasions, the clinicians scored within ±1 point of the expert, and on at least two-thirds of the occasions, clinician-expert rater pairs were fully concordant. For the measures of functional capacity (specifically the distance walked in 2 minutes and grip strength), more than 60% of the time, the clinicians and expert were in complete accord, as defined by the 95% CI interval around the expert’s value for 2-minute walk distance and grip strength. Relative to global ratings of cGVHD severity, cGVHD symptom severity, and evolution of cGVHD across time, 50% to 75% of the pair-wise assessments of cGVHD severity in the final two trials were fully concordant. In globally rating cGVHD on a 7-point scale as better, worse, or stable over the past month, 40% to 63% of the ratings by the clinician-expert pairs were fully concordant, and 80% to 100% of the ratings were within ±1 point. However, in rating cGVHD symptom severity on a 0–10 scale, agreement between clinician and expert was moderate.

Discussion

This study examined the feasibility and inter-rater reliability of the NIH cGVHD response criteria when used to evaluate adults and children with a range of cGVHD manifestations in four successive pilot trials. We observed that use of the NIH response criteria is feasible for clinicians and patients because the median time taken to complete the evaluation is appropriate for what might be expected for the complex care visits that typically occur periodically within the context of a clinical trial. However, a single training session is insufficient to achieve consistently acceptable inter-rater agreement. With refinement of the measures (clarification of the definitions of movable and nonmovable sclerosis, modification of the scale for grading oral manifestations) and enhancements to the training materials, agreement between clinicians and experts in rating selected domains of response improved across the four successive trials. For cutaneous erythema, nonmovable sclerosis, and oral manifestations, agreement between clinician-expert pairs approached what are considered fair levels of inter-rater reliability for a new measure[10]. Based on our findings, a clinical trial participant would need to demonstrate a change from baseline in the range of 19% to 22%

BSA for erythema, 18% to 26% BSA for movable sclerosis, 17% to 21% BSA for nonmovable sclerosis, and 2.1 to 2.6 points on the 15 point NIH Oral cGHVD scale, in order to conclude with 95% confidence that the changes were not attributable to variability in scoring when different raters perform the evaluation. Of note, the MDC_{95} for the 15 point NIH Oral cGVHD scale falls well within the 4 point absolute minimum change criterion proposed to establish partial response[3]. These conclusions must be tempered by the fact that in the final two trials raters had transplant experience, while earlier trials included raters with no or only limited transplant experience. It is possible that both the inclusion of raters with transplant experience (but who did not have experience with the NIH criteria), as well as the improved teaching materials jointly accounted for the improved reliability estimates observed in the final two trials.

That we observed high agreement in scoring gastrointestinal symptoms was not unexpected since these symptoms are not observed by clinicians and must be elicited directly from patients[15], and in addition, the prevalence of gastrointestinal complaints in our sample was generally low. Consistent with the findings of prior research demonstrating that patient self-report is the most valid method to evaluate symptom severity and that health care providers misestimate symptom intensities[16, 17], clinician-expert pairs were only moderately concordant in evaluating cGVHD global symptom severity on a 10 point scale. The modest level of concordance observed between experts and clinicians in rating cGVHD as better, worse, or stable on a 7-point scale may represent an artifact of our study design since the experts and the clinicians each evaluated study participants at only one time point. Thus, it must be assumed that the raters derived their scores through discussion with the patient or based on speculation. It is plausible that this feature of our study design introduced measurement error that would not be observed with serial evaluations. Future research conducted with raters who have direct knowledge of subjects' clinical status over time is warranted to estimate the reliability of the 7-point global evaluations of chronic GVHD as better, worse, or stable.

Clinician participants offered a generally favorable evaluation of the training session, opportunities for skills practice, documentation tools, and the photo atlas. The challenges in incorporating these response assessments into practice were acknowledged by participants. However it should be noted that the NIH response assessments have been developed as outcome measures in clinical trials and were not intended for routine use in clinical care. While it required a median of 21 minutes for the clinician to conduct and document the entire response evaluation, the Schirmer tear test was the most time consuming single component of the evaluation, requiring a median of 9 minutes. Therefore, the response evaluation time could be shortened to approximately 10 minutes at many centers where the evaluation of the Schirmer tear test, grip strength, and 2-minute walk test can be delegated to other staff, thus conserving clinician effort during busy clinics.

In the final two trials, the only ICC below 0.40 was observed for the evaluation of movable sclerotic features (ICC=0.24), and examination of the Bland-Altman plots suggested a tendency for the novice clinician raters to score a greater extent of involvement with movable sclerosis relative to the experts. There are several possible reasons why the measurement of movable sclerosis poses particular difficulty, as compared to the measurement of erythema or nonmovable sclerosis. In some cases, movable sclerotic cGVHD manifestations may be subtle and difficult to distinguish from non-sclerotic changes. Gauging movable sclerotic manifestations is also a complex determination because the evaluator is required to integrate visual and tactile assessments to formulate the score. In addition, the margins of movable sclerotic involvement can be challenging to define, while in contrast, in patients with nonmovable sclerosis/hidebound manifestations, differentiating affected and unaffected areas is perhaps less subtle. Furthermore, in some cases, movable cutaneous changes, particularly in the setting of other manifestations such as edema may mask underlying nonmovable sclerosis, making it difficult to distinguish the percent BSA associated with each type of involvement. Our observation that novice clinician raters in the final trial tended to overestimate movable sclerosis and underestimate nonmovable sclerosis is consistent with this possibility. If this finding is replicated in larger samples, clearer definitions to help distinguish between movable and nonmovable sclerosis should be considered to achieve greater reliability between raters. Future research to develop digitized, computer analyzed methods for estimating percent of BSA involved with specific cutaneous manifestations[18, 19] may also contribute to improvements in making such distinctions.

Other approaches to reducing measurement error include opportunities for practice, careful calibration among raters, improvements in the design of assessment tools, and using an average of multiple independent ratings rather than a single rating[20, 21]. A more precise operational definition of what constitutes movable sclerosis and improvements to the photo atlas and training materials may also improve the inter-rater reliability between clinicians and experts. For example, revision of the definitions of movable sclerosis to incorporate textural changes might increase clinicians' precision in recognizing and measuring this subtle manifestation.

Determining whether change has occurred in response to therapy requires reliable measures of therapeutic response. Two prior studies have specifically addressed the inter-rater reliability of a component of therapeutic response in cGVHD[22, 23]. Our intraclass correlation coefficients for the cutaneous manifestations of erythema and nonmovable sclerosis are comparable to those estimates reported by Greinix et al, in their evaluation of a scoring system focused on the cutaneous manifestations of cGVHD[22]. However the Greinix et al. study examined the concordance among four raters, only one of whom was described as an expert, whereas we report pair-wise agreement between experts and 34 clinicians across a wide range of patients. Treister et al. assessed the inter-rater reliability among transplant clinicians who evaluated oral cGVHD manifestations through the use of photos. Although the intrarater reliability of assessments performed one week apart was excellent, inter-rater reliability of assessments based on the photos was poor to moderate (0.15 to 0.46)[23]. Half of the participants in the Treister et al. study noted that training would have improved their accuracy in evaluating oral cGVHD manifestations[23], and the incremental improvement in reliability observed in our study suggests that even brief training strengthens clinicians' oral cGVHD assessment skills. The range of intraclass correlation coefficients observed in our study also parallels values that have been reported for the psychometric evaluation of clinical examination methods[24] and scoring systems for grading dermatologic, oral, and rheumatologic disease manifestations[25–30].

With replication in a larger sample, our preliminary estimates of the inter-rater reliability (ICC) and thus the minimum metrically detectable change that exceeds measurement error, may be helpful in future studies for power analysis and sample size estimation, and to interpret the results of prospective clinical trials [31, 32]. For example, if a new antifibrotic agent reduces the extent of cutaneous sclerosis from a mean of 60% BSA involvement to a mean of 30% BSA involvement in the sample, knowing that the MDC_{95} is 17% BSA, permits a conclusion that there has been a true therapeutic effect. The minimum metrically detectable change is distinct from, but complementary to definitions of clinically significant response (partial and complete responses) [11, 12]. Estimates of the MDC_{95} could be particularly useful in detecting therapeutic activity in early phase clinical trials of new agents. Prospective studies would then be needed to determine the magnitude of change that constitutes a clinically meaningful response.

Within-rater variation may partially account for some of the variability we observed between raters. Future studies should address this limitation by incorporating a more comprehensive design that considers both within-observer and between-observer variation, including the variation within and between expert raters. Until further reliability studies have been conducted in larger samples of children with cGVHD, caution should be exercised when applying these inter-rater reliability estimates in pediatric clinical trials. We recognize that the sample size for this series of pilot trials was also relatively small, and clinician raters received just a single 2.5 hour training session with no opportunity to gain experience in applying the response criteria before conducting these one-time response assessments on patients. Raters may have been able to improve in their skill and accuracy with more opportunity to practice these new skills on several patients prior to obtaining the ratings used to calculate these reliability estimates. However, a strength of this study is that the inter-rater reliability estimates were obtained in a naturalistic setting, and involved study participants exhibiting a wide range of cGVHD manifestations. Moreover, to our knowledge, this is the first study to test the feasibility and reproducibility of a multidimensional set of cGVHD therapeutic response criteria, and to evaluate the response criteria in both adults and children.

While these findings support a conclusion that application of the NIH criteria in determining therapeutic response will contribute to uniformity of data collection and advance standards in cGVHD clinical trials, our findings point to several important directions for future research. There is a need to evaluate inter- and intra-rater reliability of the NIH cGVHD response criteria in prospective studies, and to examine the concurrent and predictive validity of the response criteria. Evaluation of inter-rater reliability in a larger sample of raters would permit the application of multivariate techniques such as receiver operating characteristic (ROC) analysis and would help to overcome the constraints inherent when generalizing estimates of inter-rater agreement developed on small samples of observers[33]. Composite elements in the response criteria should be individually evaluated for their reliability, validity, and sensitivity to change. For example, studies examining the psychometric properties such as the test-retest reliability and sensitivity to change of the 2 minute walk distance and grip strength in the cGVHD population will contribute to greater precision in our estimates of reproducibility and will assist in defining clinically meaningful change. Similarly, research is warranted to compare the performance of methods to quantify cGVHD ocular disease manifestations (e.g. Schirmer tear test, tear film breakup time, tear osmolarity)[34], and to identify the method(s) with greatest reliability, specificity, and sensitivity to change. Research to determine the effect that raters' level of transplant and cGVHD experience has on the achievement of sufficiently reliable scoring is needed. Future studies are also indicated to determine if more intensive training that includes repetition, calibration, and feedback on performance from experts in cGVHD cutaneous manifestations, and the use of standardized patients and simulation, will improve the accuracy with which cutaneous manifestations are recognized and scored. Lastly, the evaluation of cGVHD therapeutic response is an area that would benefit from efforts to develop intermediate endpoints of response including biomarkers, and to automate and digitize assessments through refined instrumentation, imaging methods, and computer-based applications[19, 35–38], thus improving reliability by limiting potential sources of measurement error.

Acknowledgments

We gratefully acknowledge Olena Prachenko, MA and Leanne Silverman, BA, for assistance with data analysis and manuscript preparation, and Nancy Kline Leidy, PhD, United Biosource Corporation, and Gwenyth R. Wallen, RN, PhD, National Institutes of Health, for valuable critique. This work was supported by the Intramural Research Program at the National Institutes of Health Clinical Center, and by the Center for Cancer Research, National Cancer Institute.

References

1. Socie G, Ritz J, Martin PJ. Current challenges in chronic graft-versus-host disease. Biol Blood Marrow Transplant. Jan; 2010 16(1 Suppl):S146–151. [PubMed: 19836455]
2. Martin PJ, Weisdorf D, Przepiorka D, et al. National Institutes of Health Consensus Development Project on Criteria for Clinical Trials in Chronic Graft-versus-Host Disease: VI. Design of Clinical Trials Working Group report. Biol Blood Marrow Transplant. May; 2006 12(5):491–505. [PubMed: 16635784]
3. Pavletic SZ, Martin P, Lee SJ, et al. Measuring therapeutic response in chronic graft-versus-host disease: National Institutes of Health consensus development project on criteria for clinical trials in chronic graft-versus-host disease: IV. Response criteria working group report. Biol Blood Marrow Transplant. Mar; 2006 12(3):252–266. [PubMed: 16503494]
4. Lee S, Cook EF, Soiffer R, Antin JH. Development and validation of a scale to measure symptoms of chronic graft-versus-host disease. Biol Blood Marrow Transplant. 2002; 8(8):444–452. [PubMed: 12234170]
5. Sanchez MM, Binkowitz BS. Guidelines for measurement validation in clinical trial design. J Biopharm Stat. Aug; 1999 9(3):417–438. [PubMed: 10473029]
6. Shrout P, Fleiss J. Intraclass correlations: Uses in assessing rater reliability. Psychological Bulletin. 1979; 2:420–428. [PubMed: 18839484]
7. Shoukri, MM. Measures of interobserver agreement. Boca Raton, FL: Chapman & Hall/CRC; 2004.
8. Denegar CR, Ball DW. Assessing Reliability and Precision of Measurement: An Introduction to Intraclass Correlation and Standard Error of Measurement. Journal of Sport Rehabilitation. 1993; 2:35–42.
9. McGraw KL, Wong SP. Forming inferences about some intraclass correlation coefficients. Psychological Methods. 1996; 1(1):30–46.
10. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. Mar; 1977 33(1):159–174. [PubMed: 843571]
11. Beaton DE. Understanding the relevance of measured change through studies of responsiveness. Spine. Dec 15; 2000 25(24):3192–3199. [PubMed: 11124736]
12. Beaton DE, Bombardier C, Katz JN, Wright JG. A taxonomy for responsiveness. J Clin Epidemiol. Dec; 2001 54(12):1204–1217. [PubMed: 11750189]
13. Jacobsohn DA, Rademaker A, Kaup M, Vogelsang GB. Skin response using NIH consensus criteria vs Hopkins scale in a phase II study for steroid-refractory chronic GVHD. Bone Marrow Transplant. Dec; 2009 44(12):813–819. [PubMed: 19430498]
14. Filipovich AH, Weisdorf D, Pavletic S, et al. National Institutes of Health consensus development project on criteria for clinical trials in chronic graft-versus-host disease: I. Diagnosis and staging working group report. Biol Blood Marrow Transplant. Dec; 2005 11(12):945–956. [PubMed: 16338616]
15. Basch E, Iasonos A, McDonough T, et al. Patient versus clinician symptom reporting using the National Cancer Institute Common Terminology Criteria for Adverse Events: results of a questionnaire-based study. Lancet Oncol. Nov; 2006 7(11):903–909. [PubMed: 17081915]
16. Justice AC, Chang CH, Rabeneck L, Zackin R. Clinical importance of provider-reported HIV symptoms compared with patient-report. Medical Care. Apr; 2001 39(4):397–408. [PubMed: 11329526]
17. Laugsand EA, Sprangers MA, Bjordal K, Skorpen F, Kaasa S, Klepstad P. Health care providers underestimate symptom intensities of cancer patients: a multicenter European study. Health Qual Life Outcomes. 2010; 8:104. [PubMed: 20858248]
18. Bae Y, Son T, Stuart Nelson J, Kim JH, Choi EH, Jung B. Dermatological feasibility of multimodal facial color imaging modality for cross-evaluation of facial actinic keratosis. Skin Res Technol. Feb; 2011 17(1):4–10. [PubMed: 20923462]
19. Zulian F, Meneghesso D, Grisan E, et al. A new computerized method for the assessment of skin lesions in localized scleroderma. Rheumatology (Oxford). May; 2007 46(5):856–860. [PubMed: 17264088]

20. Czirjak L, Nagy Z, Aringer M, Riemekasten G, Matucci-Cerinic M, Furst DE. The EUSTAR model for teaching and implementing the modified Rodnan skin score in systemic sclerosis. Ann Rheum Dis. Jul; 2007 66(7):966–969. [PubMed: 17234649]

21. Leon AC, Marzuk PM, Portera L. More reliable outcome measures can reduce sample size requirements. Arch Gen Psychiatry. Oct; 1995 52(10):867–871. [PubMed: 7575107]

22. Greinix HT, Pohlreich D, Maalouf J, et al. A single-center pilot validation study of a new chronic GVHD skin scoring system. Biol Blood Marrow Transplant. Jun; 2007 13(6):715–723. [PubMed: 17531782]

23. Treister NS, Stevenson K, Kim H, Woo SB, Soiffer R, Cutler C. Oral chronic graft-versus-host disease scoring using the NIH consensus criteria. Biol Blood Marrow Transplant. Jan; 2010 16(1): 108–114. [PubMed: 19772943]

24. Koran LM. The reliability of clinical methods, data and judgments (second of two parts). N Engl J Med. Oct 2; 1975 293(14):695–701. [PubMed: 1160937]

25. Gladman DD, Inman RD, Cook RJ, et al. International spondyloarthritis interobserver reliability exercise--the INSPIRE study: II. Assessment of peripheral joints, enthesitis, and dactylitis. J Rheumatol. Aug; 2007 34(8):1740–1745. [PubMed: 17659754]

26. Aktan S, Ilknur T, Akin C, Ozkan S. Interobserver reliability of the Nail Psoriasis Severity Index. Clin Exp Dermatol. Mar; 2007 32(2):141–144. [PubMed: 17137477]

27. Arkachaisri T, Pino S. Localized scleroderma severity index and global assessments: a pilot study of outcome instruments. J Rheumatol. Apr; 2008 35(4):650–657. [PubMed: 18322985]

28. Bhakta BB, Brennan P, James TE, Chamberlain MA, Noble BA, Silman AJ. Behcet's disease: evaluation of a new instrument to measure clinical activity. Rheumatology (Oxford). Aug; 1999 38(8):728–733. [PubMed: 10501420]

29. Mulic A, Tveit AB, Wang NJ, Hove LH, Espelid I, Skaare AB. Reliability of two clinical scoring systems for dental erosive wear. Caries Res. Jul; 2010 44(3):294–299. [PubMed: 20516691]

30. Ionescu R, Rednic S, Damjanov N, et al. Repeated teaching courses of the modified Rodnan skin score in systemic sclerosis. Clin Exp Rheumatol. Mar–Apr; 2010 28(2 Suppl 58):S37–41. [PubMed: 20576212]

31. Perkins DO, Wyatt RJ, Bartko JJ. Penny-wise and pound-foolish: the impact of measurement error on sample size requirements in clinical trials. Biol Psychiatry. Apr 15; 2000 47(8):762–766. [PubMed: 10773186]

32. Muller MJ, Szegedi A. Effects of interrater reliability of psychopathologic assessment on power and sample size calculations in clinical trials. J Clin Psychopharmacol. Jun; 2002 22(3):318–325. [PubMed: 12006903]

33. Sadatsafavi M, Najafzadeh M, Lynd L, Marra C. Reliability studies of diagnostic tests are not using enough observers for robust estimation of interobserver agreement: a simulation study. J Clin Epidemiol. Jul; 2008 61(7):722–727. [PubMed: 18486446]

34. Versura P, Profazio V, Campos EC. Performance of tear osmolarity compared to previous diagnostic tests for dry eye diseases. Curr Eye Res. Jul; 35(7):553–564. [PubMed: 20597641]

35. Moore TL, Lunt M, McManus B, Anderson ME, Herrick AL. Seventeen-point dermal ultrasound scoring system--a reliable measure of skin thickness in patients with systemic sclerosis. Rheumatology (Oxford). Dec; 2003 42(12):1559–1563. [PubMed: 12867579]

36. Kuwahara Y, Shima Y, Shirayama D, et al. Quantification of hardness, elasticity and viscosity of the skin of patients with systemic sclerosis using a novel sensing device (Vesmeter): a proposal for a new outcome measurement procedure. Rheumatology (Oxford). Jul; 2008 47(7):1018–1024. [PubMed: 18440998]

37. Merkel PA, Silliman NP, Denton CP, et al. Validity, reliability, and feasibility of durometer measurements of scleroderma skin disease in a multicenter treatment trial. Arthritis Rheum. May 15; 2008 59(5):699–705. [PubMed: 18438905]

38. Sato LT, Kayser C, Andrade LE. Nailfold capillaroscopy abnormalities correlate with cutaneous and visceral involvement in systemic sclerosis patients. Acta Reumatol Port. Apr–Jun; 2009 34(2A):219–227. [PubMed: 19474777]

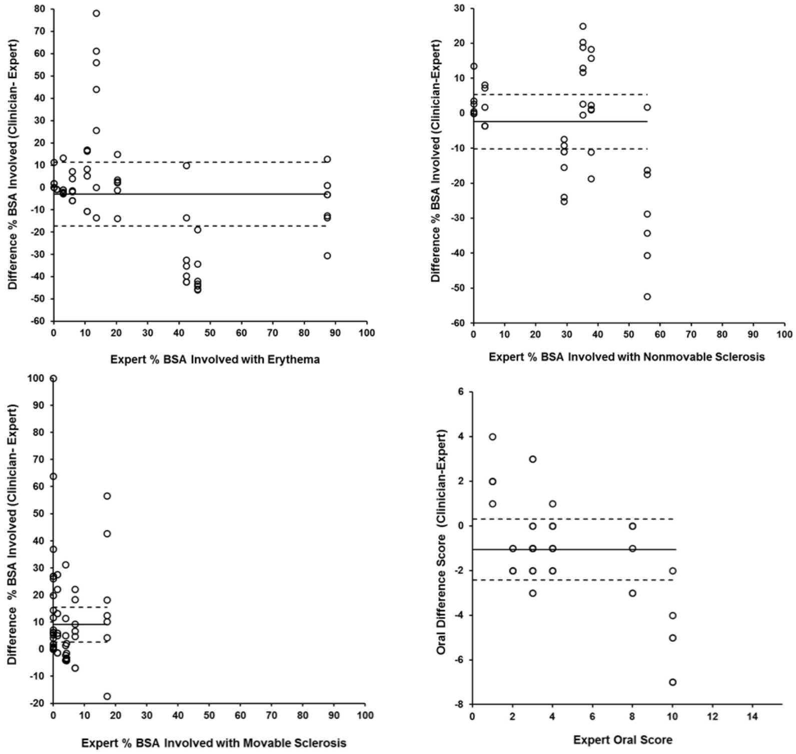

Figure 1.

Bland-Altman plots, comparing differences (Trial 4) between clinician and expert scoring of cutaneous and oral cGVHD manifestations (difference=clinician score minus expert score) plotted against expert scores. Negative differences reflect clinician underestimation of the extent of involvement, while positive differences reflect clinician overestimation of the extent of involvement, relative to the expert’s assessment. A change in the magnitude of the difference between clinician and expert assessments with increasing extent of cGVHD involvement, as assessed by the expert, is determined by looking for patterns along the x-axis.

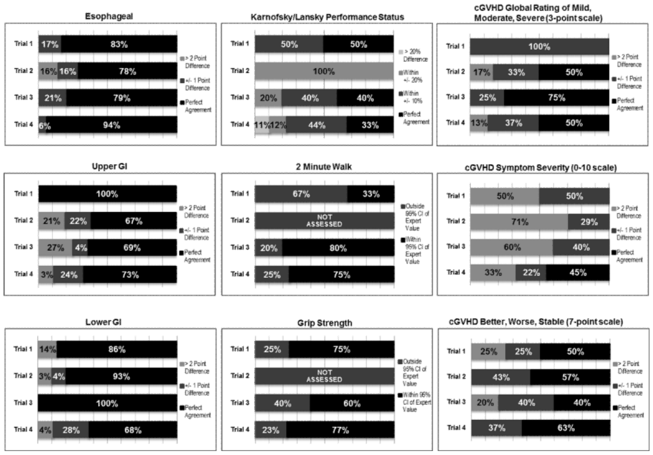

Figure 2.

Inter-rater Agreement Between Clinicians and Experts for Evaluation of Gastrointestinal Symptoms, Functional Performance, and cGVHD Global Scores. In the final two trials, substantial inter-rater agreement was observed in evaluating gastrointestinal symptoms (68% to 94% of pair-wise comparisons in perfect agreement), while moderate to substantial agreement was noted in evaluating the two minute walk (75% to 80% of pair-wise assessments were concordant), grip strength (60% to 77% of pair-wise assessments were concordant), cGVHD global severity (clinician-expert pairs in perfect agreement 50% to 75% of the time), and cGVHD evolution over the past month (clinician-expert pairs in perfect agreement 40% to 63% of the time).

Note: Numbers were rounded so that values add to 100%

Table 1

Chronic GVHD Characteristics in the Sample

	Overall (N=25)	Trial 1 (N=4)	Trial 2 (N=6)	Trial 3 (N=5)	Trial 4 (N=10)
Cutaneous CGVHD Body Surface Area (BSA) Involved					
Erythema (Mean %BSA/S.D.)	20.9 (26.5)	30.0 (47.6)	19.3 (23.2)	11.9 (15.8)	22.9 (27.8)
Movable Sclerosis (Mean %BSA/S.D.)	14.1 (15.6)	17.5 (17.1)	16.9 (17.8)	22.7 (20.9)	3.4 (5.5)
Nonmovable Sclerosis (Mean %BSA/S.D.)	13.4 (18.3)	14.5 (23.7)	3.7 (4.9)	9.6 (12.6)	16.1 (21.1)
Presence of Ulcers	6 (24%)	1 (25%)	2 (33%)	3 (60%)	0 (0%)
NIH Oral cGHVD scale (0–15 point scale (Mean/S.D.)	4.2 (2.1)	5.3 (2.4)	3.5 (0.6)	4.4 (1.7)	4.1 (2.8)
Schirmer's Tear Test (mm of wetting) (Mean/S.D.)	10.6 (8.9)	5.9 (11.4)	8.9 (6.1)	11.2 (9.8)	13.4 (8.7)
2 Minute Walk Distance (total feet/2 minutes) (Mean/S.D.)	563 (147.3)	655 (63.5)	367 (173.4)	598 (66.3)	502(182.3)
Grip Strength (psi)(Mean/S.D.)	50.8 (32.1)	59.8 (31.9)	23.8 (20.0)	68.2 (33.3)	37.2 (28.5)
Lee CGVHD Symptom Bother (0–100 point scale; higher scores reflect greater cGVHD symptom bother) (Mean/S.D.)	18.9 (13.4)	22.3 (12.9)	24.9 (21.1)	14.8 (10.5)	17.5 (13.7)
Upper GI Symptoms (Early satiety, anorexia, nausea/vomiting)					
None	21 (84%)	4 (100%)	4 (67%)	4 (80%)	9 (90%)
Mild, occasional symptoms	2 (8%)	0 (0%)	0 (0%)	1 (20%)	1 (10%)
Moderate, intermittent symptoms; no reduced intake	1 (4%)	0 (0%)	1 (17%)	0 (0%)	0 (0%)
Severe, persistent symptoms; marked reduction intake	1 (4%)	0 (0%)	1 (17%)	0 (0%)	0 (0%)
Esophageal Symptoms					
None	21 (84%)	4 (100%)	5 (83%)	4 (80%)	8 (80%)
Occasional dysphagia/odynophagia with solid food/pills	3 (12%)	0 (0%)	0 (0%)	1 (20%)	2 (20%)
Intermittent dysphagia/odynophagia with solid food/pills	1 (4%)	0 (0%)	1 (17%)	0 (0%)	0 (0%)
Dysphagia or odynophagia for almost all oral intake	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
Lower GI Symptoms					
None	19 (76%)	3 (75%)	4 (67%)	5 (100%)	7 (70%)
Occasional loose or liquid stools	2 (8%)	0 (0%)	0 (0%)	0 (0%)	2 (20%)
Intermittent loose or liquid stools throughout the day	4 (16%)	1 (25%)	2 (33%)	0 (0%)	1 (10%)
Voluminous diarrhea requiring intervention	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
Clinician Rating of cGVHD Symptom Severity (0–10 point scale) (Mean/S.D.)	4.9 (2.1)	4.8 (1.0)	6.5 (1.5)	5.0 (2.9)	4.1 (2.0)
Chronic GVHD NIH Severity Score					
Mild	7 (28%)	1 (25%)	0 (0%)	2 (40%)	4 (40%)

	Overall (N=25)	Trial 1 (N=4)	Trial 2 (N=6)	Trial 3 (N=5)	Trial 4 (N=10)
Moderate	8 (32%)	0 (0%)	4 (67%)	1 (20%)	3 (30%)
Severe	10 (40%)	3 (75%)	2 (33%)	2 (40%)	3 (30%)
Clinician Evaluation of cGVHD Evolution Over Past Month					
Improved	10 (40%)	3 (75%)	1 (17%)	1 (20%)	5 (50%)
Worsened	3 (12%)	0 (0%)	1 (17%)	1 (20%)	1 (10%)
Unchanged	12 (48%)	1 (25%)	4 (67%)	3 (60%)	4 (40%)

Table 2

Inter-rater Variability for Evaluation of Cutaneous Manifestations of CGVHD

	Median Intraclass Correlation Coefficient (ICC)	Absolute Difference Between Clinician and Expert %BSA Scores Mean (±S.D.)	Limits of Agreement Clinician-Expert Concordance Within ± 10% BSA N (%)	Minimum Metrically Detectable Change (MDC_{95}) in Percent BSA
Trial 1 [Adults (N=4) 28 rating tasks]				
Erythema	0.88	21.72 (9.60)	9 (32%)	10
Movable Sclerosis	0.33	16.98 (11.65)	13 (46%)	24
Nonmovable Sclerosis	0.23	19.01 (6.18)	14 (50%)	30
Trial 2 [Pediatrics (N=6) 60 rating tasks]				
Erythema	0.07	16.48 (4.55)	36 (60%)	28
Movable Sclerosis	0.21	17.04 (8.90)	30 (50%)	26
Nonmovable Sclerosis	0.16	12.27 (7.52)	43(72%)	32
Trial 3 [Adults (N=5) 44 rating tasks]				
Erythema	0.47	10.42 (4.25)	27 (61%)	22
Movable Sclerosis	0.60	11.35 (4.09)	24 (55%)	18
Nonmovable Sclerosis	0.62	10.44 (8.87)	31 (70%)	21
Trial 4 [Adults and Pediatrics (N=10) 65 rating tasks]				
Erythema	0.57	16.31 (3.10)	35 (54%)	19
Movable Sclerosis	0.24	9.12 (6.52)	47 (72%)	*26*
Nonmovable Sclerosis	0.76	7.19 (3.18)	47 (72%)	17

Table 3

Inter-rater Variability for Evaluation of Oral Manifestations of CGVHD

	Median Intraclass Correlation Coefficient (ICC)	**Absolute Difference Between Clinician and Expert Score-Mean (±S.D.)**	**Limits of Agreement Clinician-Expert Concordance within 1 point on a 15 point scale**	**Minimum Metrically Detectable Change (MDC_{95}) (15 point scale)**
Trial 1 [Adults (28 rating tasks)]				
Oral Manifestations	0.50	2.23 (0.96)	11 (39%)	2.7 points
Trial 2 [Pediatrics (60 rating tasks)]				
Oral Manifestations	0.44	1.57 (0.60)	36 (60%)	2.9 points
Trial 3 [Adults (40 rating tasks)]				
Oral Manifestations	0.57	2.71 (1.51)	17 (43%)	2.6 points
Trial 4 [Adults and Pediatrics (66 rating tasks)]				
Oral Manifestations	0.70	1.61 (0.48)	38 (58%)	2.1 points