



Published in final edited form as:

Ophthalmology. 2012 March ; 119(3): 487–493. doi:10.1016/j.ophtha.2011.08.040.

Validation of Measurement Scales in Ocular Graft-versus-Host Disease

Yoshihiro Inamoto, MD¹, Xiaoyu Chai, MS¹, Brenda F. Kurland, PhD¹, Corey Cutler, MD, MPH², Mary E.D. Flowers, MD¹, Jeanne M. Palmer, MD³, Paul A. Carpenter, MD¹, Mary J. Heffernan, ARNP¹, David Jacobsohn, MD⁴, Madan H. Jagasia, MD⁵, Joseph Pidala, MD⁶, Nandita Khera, MD¹, Georgia B. Vogelsang, MD⁷, Daniel Weisdorf, MD⁸, Paul J. Martin, MD¹, Steven Z. Pavletic, MD⁹, Stephanie J. Lee, MD, MPH¹, and Chronic GVHD Consortium

¹Division of Clinical Research, Fred Hutchinson Cancer Research Center, Seattle, Washington

²Hematologic Malignancies, Dana-Farber Cancer Institute, Boston, Massachusetts ³Division of

Hematology/Oncology, Medical College of Wisconsin, Milwaukee, Wisconsin ⁴Division of Blood

and Marrow Transplantation, Children's National Medical Center, Center for Cancer and Blood

Disorders, Washington, DC ⁵Hematology and Stem Cell Transplant Program, Vanderbilt

University Medical Center, Nashville, Tennessee. ⁶Blood and Marrow Transplantation, Moffitt

Cancer Center, Tampa, Florida. ⁷Oncology, Johns Hopkins Hospital, Baltimore, Maryland. ⁸Blood

and Marrow Transplant Program, University of Minnesota, Minneapolis, Minnesota ⁹National

Cancer Institute, Bethesda, Maryland

Abstract

Purpose—To validate measurement scales for rating ocular chronic graft-versus-host disease (GVHD) after allogeneic hematopoietic cell transplantation. Candidate scales were recommended for use in clinical trials by the National Institutes of Health (NIH) Chronic GVHD Consensus Conference or have been previously validated in dry eye syndromes.

Design—Prospective follow-up study.

Participants—Between August 2007 and June 2010, the study enrolled 387 patients with chronic GVHD in a multicenter, prospective, observational cohort.

Methods—Using anchor-based methods, we compared clinician or patient-reported changes in eye symptoms (8-point scale) with calculated changes in 5 candidate scales: The NIH eye score, patient-reported global rating of eye symptoms, Lee eye subscale, Ocular Surface Disease Index, and Schirmer test. Change was examined for 333 follow-up visits where both clinician and patient reported eye involvement at the previous visit. Linear mixed models were used to account for within-patient correlation.

Main Outcome Measures—An 8-point scale of clinician or patient-reported symptom change was used as an anchor to measure symptom changes at the follow-up visits.

Results—In serial evaluations, agreement regarding improvement, stability, or worsening between the clinician and patient was fair (weighted kappa = 0.34). Despite only fair agreement between evaluators, all scales except the Schirmer test correlated with both clinician-reported and

patient-reported changes in ocular GVHD activity. Among all scales, changes in the NIH eye scores showed the greatest sensitivity to symptom change reported by clinicians or patients.

Conclusions—Our results support the use of the NIH eye score as a sensitive measure of eye symptom changes in clinical trials assessing treatment of chronic GVHD.

Allogeneic hematopoietic cell transplantation (HCT) is a curative treatment for many hematologic malignancies and nonmalignancies,¹ but chronic graft-versus-host disease (GVHD) is the leading cause of late morbidity in transplant survivors, compromising both quality of life and function.²⁻⁸ Chronic GVHD is thought to occur because the donor's immune system recognizes recipient tissues, causing inflammation and fibrosis. Eye involvement occurs in 40% to 60% of patients with chronic GVHD,⁹⁻¹¹ and ocular symptoms may be the presenting feature of chronic GVHD. Many patients need ancillary and supportive care for the eye, including artificial tears or other ocular lubricants, special eyewear, or topical immunosuppressive agents, in addition to systemic immunosuppressive treatment.^{12,13}

Clinical trials in chronic GVHD have been hampered by lack of validated clinical or biologic measures that can capture the severity and response to treatment. In 2005, the National Institutes of Health (NIH) convened a consensus conference on criteria for clinical trials in chronic GVHD to propose such measures. To be considered for inclusion in the recommendations, a measure had to be easy to use by both transplantation and nontransplantation care providers and easily available in the outpatient setting. For this reason, special ophthalmologic examinations such as slit-lamp examination and corneal staining are not included in the recommendation for eye assessment. The NIH consensus conference distinguished between “severity” measures, which were considered cross-sectional assessment tools that reflect current chronic GVHD involvement,¹⁴ and “response” measures, which were intended to be assessed serially and thought to be more sensitive to changes owing to effective treatments.¹⁵ For example, the NIH severity assessment for the eye is based on 4 categories that physicians report.¹⁴ In contrast, the NIH-recommended tools for response assessment included both the Schirmer test and patient-reported outcomes,^{16,17} such as the patient-reported global rating of eye symptom¹⁸ and the Lee symptom scale.¹⁹ In addition to all the recommended scales, we also tested the Ocular Surface Disease Index (OSDI), although this scale was originally designed to measure the severity of ocular manifestations of Sjögren's syndrome.²⁰

To validate the measures proposed by the NIH consensus conference and select the best among them, we compared clinician or patient-reported eye symptom changes with calculated changes in serial measurement scales using a prospective cohort of patients with chronic GVHD. We used anchor-based methods to identify the measurement scales that were most sensitive to reported symptom change (8-point scale).²¹ Anchor-based methods examine the relationship between scores on the instrument whose interpretation is under question (target instrument) and some independent measure (an “anchor,” the gold standard). Our goal was to eliminate insensitive or duplicative measurement scales and recommend a parsimonious battery of measurement scales for use in clinical trials. This is the first of a series of organ-specific analyses to validate the NIH consensus tools, which will potentially form the basis of revised NIH recommendations.

Patients and Methods

Patients were enrolled in a multicenter, prospective, observational cohort in which all the measures recommended by the NIH chronic GVHD consensus conference were collected. The rationale and design of the cohort study is described in detail elsewhere.²² One of the specific aims of this cohort study was to validate the various measures for the different

organ systems. The study protocol was approved by the institutional review board of each participating center, and all participants or their guardians gave written informed consent. This analysis examined data collected through June 2010.

Patient Eligibility and Follow-up

Diagnosis of chronic GVHD was made clinically according to the NIH consensus criteria.¹⁴ Distinctive manifestations of ocular GVHD include new onset of dry, gritty, or painful eyes; cicatricial conjunctivitis; keratoconjunctivitis sicca; and confluent areas of punctate keratopathy. New ocular sicca documented by low Schirmer scores, or a new onset of keratoconjunctivitis sicca by slit-lamp examination is sufficient for the diagnosis if accompanied by distinctive manifestations of chronic GVHD in another organ. Patients ≥ 2 years of age with chronic GVHD requiring systemic treatment were enrolled. Systemic treatment was defined as any medication or intervention delivered systemically, including extracorporeal photopheresis. Enrollment included both incident cases (enrollment < 3 months after chronic GVHD diagnosis) and prevalent cases (enrollment ≥ 3 months after chronic GVHD diagnosis, but within 3 years after HCT). Allogeneic HCT for any disease with any graft source, donor type, and GVHD prophylaxis was allowed. Patients with recurrent disease at enrollment or anticipated survival of < 6 months were excluded.

Patients were evaluated at the transplant center every 6 months. Incident cases had an additional assessment 3 months after enrollment. Patients were treated according to institutional practice, although compliance with the NIH chronic GVHD consensus guidelines for supportive care was encouraged.^{12,15}

Measurement Scales and Reported Symptom Change

Validated measurement scales included the NIH eye score reported by clinicians¹⁴ (1 item; score range, 0–3), patient-reported global rating of eye symptoms^{15,18} (1 item; score range, 0–10), patient-reported eye subscale of Lee symptom scale¹⁹ (3 items; score range, 0–100), patient-reported OSDI²³ (12 items; score range, 0–100), and the measured Schirmer test (Fig 1). It was recommended that the Schirmer test be performed without anesthesia in adults. The value of the worst eye was used for analysis of the Schirmer test because the results were similar when the average of both eyes was used. Symptoms were assessed without a distinction of left and right eyes for other measurement scales. An 8-point scale of clinician- or patient-reported symptom change²⁴ was used as an “anchor” to measure symptom changes at the follow-up visits that ranged from 1 (completely gone) to 8 (very much worse; Fig 1).

Statistical Analysis

Clinician-reported eye involvement was defined as any symptoms by the NIH eye score. Patient-reported eye involvement was defined as any symptoms by global rating scale or the Lee eye subscale. Agreement between clinician and patient ratings of eye involvement at enrollment was tested by the kappa statistic. Empirical interpretation was used for kappa coefficient (0, no agreement; 0–0.2, slight agreement; 0.2–0.4 fair agreement; 0.4–0.6, moderate agreement; 0.6–0.8, substantial agreement; and 0.8–1.0, almost perfect agreement).

Changes in scores for validated measurement scales were calculated by subtracting the values recorded at sequential visits. Only paired visits with eye involvement reported by both clinician and patient at the previous visit were included in the longitudinal analysis. This conservative criterion for defining eye involvement restricts analysis to the population of interest, that is, patients who would be candidates for a clinical trial of eye-directed therapy because both the clinician and the patient agree that eye symptoms are present. Serial changes in Schirmer scores were analyzed using raw calculated values, and the results

were confirmed with 2 additional models exploring different definitions of serial changes, based on NIH response criteria that consider values >10 mm as normal.²⁵ In 1 model, all values >10 mm were capped at 10 mm. In another model, the serial change was reset to zero if both the initial and follow-up Schirmer scores were ≥ 10 mm.

A clinically meaningful change is the degree of change in a scale that is perceptible to the evaluator as either improvement or worsening. Cutoffs for clinically meaningful change for each scale were derived from prior publications or as recommended by the NIH consensus project. A 1-point change on a 3- or 7-point scale or 2-point change on a 10-point scale was used for global ratings and categorical scales.^{15,24} A half-standard deviation change was used for other patient-reported measures.^{15,26,27} A 5-mm change was used for the Schirmer test.²⁵

Agreement between clinician- and patient-reported eye symptom changes was tested by the kappa statistic. To measure agreement, the 8-level reported symptom change scale was collapsed into 3 categories: Improvement (1 [completely gone], 2 [very much better], 3 [moderately better]); stable (4 [a little better], 5 [about the same], 6 [a little worse]); or worsening (7 [moderately worse], 8 [very much worse]). Graphical displays were used to explore the actual changes in serial measurement scales according to clinician- or patient-reported symptom changes. Multivariable models were constructed to examine correlations of changes in measurement scales with symptom changes reported by clinicians or patients, after controlling for patient and disease characteristics. Linear mixed models with random patient effects were used to account for within-patient correlation^{28,29} because these models were little affected by missing data and the results were similar with generalized estimating equations. In multivariable models, co-variables considered were patient age at enrollment, months from HCT to enrollment, months from onset of chronic GVHD to enrollment, patient gender, status at enrollment (incident versus prevalent), diagnosis at transplant, graft type, donor-patient gender combination, human leukocyte antigen and donor type (matched sibling donor versus other), conditioning regimen (myeloablative versus other), history of acute GVHD, other organ involvement, and education level (6-level scale). The status at enrollment was included as a covariate in all multivariable models because it showed a trend to affect the results. Statistical analyses were performed with SAS/STAT software, version 9.2 (SAS Inc., Cary, NC) and R version 2.9.2 (R Foundation for Statistical Computing, Vienna, Austria).

Results

Patient Characteristics

Between August 2007 and June 2010, the study enrolled 387 patients, including 14 children (4%). The OSDI and Schirmer tests were not evaluated in children. The median patient age at enrollment was 51 years (range, 2-79 years). The median time from transplant to enrollment was 12.3 months (range, 2.9-39.2 months). The cohort included 209 incident cases (54%) and 178 prevalent cases (46%). Other demographic characteristics of patients are summarized in Table 1 (available at <http://aaojournal.org>).

Agreement in Clinician and Patient Ratings of Eye Involvement and Change in Ocular GVHD Activity

Fifty-two of 387 cases (13%) lacked eye measurement scales at enrollment. Among the remaining 335 cases at enrollment, eye involvement was reported by clinicians in 175 cases (52%), and by patients in 231 cases (69%) (kappa = 0.40, fair agreement; Fig 2A). Both the clinician and patient reported eye involvement at the previous visit in 333 of the 763 follow-up visits (44%). Fifty-six of these 333 visits (17%) were missing either clinician-

or patient-reported symptom change measures. For the remaining 277 visits (Fig 2B), both the clinician and patient most often reported symptom change as stable (66% and 52%, respectively), with improvement (30% and 35%, respectively) more likely than worsening (3% and 12%, respectively). The agreement regarding improvement, stability, or worsening between the clinician and patient at any given visit was fair (Fleiss-Cohen weighted kappa = 0.34).

Relationship between Reported Symptom Change and Serial Change in Validated Scales

The actual changes in measurement scales according to clinician or patient-reported symptom change are shown in Figure 3 (available at <http://aojournal.org>). The NIH eye score, the global rating of eye symptoms, the Lee eye subscale, and the OSDI seemed to agree with both clinician and patient-reported symptom change (Fig 3A-D and F-I). Changes in Schirmer scores agreed with clinician-reported symptom change (Fig 3E), but showed less correlation with patient-reported symptom change (Fig 3J).

The proportions of visits with clinically meaningful improvement or worsening were similar for all 5 scales (Table 2), but agreement between the scales was no better than moderate (Fig 4). The results of multivariable models are shown in Table 2. Changes in all scales except the Schirmer test were statistically correlated with both clinician- and patient-reported symptom changes. Changes in Schirmer scores were not correlated with patient-reported symptom changes. Among the scales validated, the magnitude of estimated change in reported symptom changes was greatest for the NIH eye score both in clinician- and patient-reported symptom changes (point estimates of reported symptom change by clinicians and patients for the NIH eye score were 1.26 and 0.43, respectively). The results for the Schirmer test were confirmed with 2 additional models (see Methods). Both models showed results similar to the original results (data not shown). The results of Schirmer tests were missing in 123 of 333 follow-up visits (37%) and serial changes in Schirmer scores were not available in >50% of the follow-up visits on longitudinal analysis. Missing Schirmer tests were correlated with higher NIH eye score ($P = 0.02$), but not with patient age, patient gender, other scale scores, or previous Schirmer scores.

Discussion

We tested 3 recommended measures for response in ocular GVHD (the Schirmer test, patient-reported global rating of eye symptoms, and the Lee eye subscale), 1 experimental measure (the OSDI), and the NIH eye score. Our results suggested strong correlations of serial changes in all measurement scales with clinician- and patient-reported symptom changes, except for the Schirmer test. The sensitivity of each scale to symptom changes reported by clinicians or patients varied among the scales.

The NIH eye score was originally intended to evaluate the severity of ocular GVHD at single visits, but our results support the conclusion that changes in the NIH eye score between visits could also be used to evaluate response. In fact, the NIH eye score seemed to be the most sensitive to symptom changes reported by both the clinician and the patient, because a change of 1 unit in the 4-level NIH eye score was associated with the greatest estimate of symptom change reported by clinicians and patients among the validated scales. From a logistic point of view, the NIH eye score required less time to collect than the other measurement scales, was collected more frequently, and is therefore most likely to be captured reliably and efficiently.

The patient-reported global rating of eye symptoms includes the “chief eye complaint” rated in the 10-point scale for peak severity during the past week.^{15,18} A major advantage of this measurement scale is the minimal time and effort for patients to complete it. A criticism has

been that this scale might reflect ocular symptoms unrelated to GVHD, but this concern is not supported by our results; changes in this global rating scale correlated with clinician-reported symptom changes in ocular GVHD.

The Lee symptom scale is a patient self-administered questionnaire consisting of 30 items specific to symptoms of chronic GVHD.¹⁹ Three of the 30 items refer to eye symptoms (Fig 1). Changes in the Lee eye subscale correlated with both clinician- and patient-reported symptom changes in ocular GVHD. Although items appeared more specific to symptoms of ocular GVHD than the global rating of eye symptoms, the magnitude of estimated change in reported symptoms was very similar between the 2 NIH-recommended patient-reported measurement scales.

The OSDI is a valid and reliable scale for diagnosis and measurement of the severity of dry eye disease.²³ Schiffman et al²³ recommended the use of OSDI as an endpoint in clinical trials. Several recent studies of dry eye syndrome³⁰⁻³² and 1 study of ocular GVHD³³ used changes in OSDI as an endpoint. Our results showed that changes in OSDI correlated with clinician- and patient-reported symptom changes in ocular GVHD. The magnitudes of estimated change in reported symptoms were comparable among the global rating scale, the Lee eye subscale, and the OSDI (Table 2). Given that the OSDI consists of 12 items instead of the 3 for the Lee eye subscale or 1 for the global rating scale, we find little basis for recommending the OSDI to measure change in activity of ocular GVHD.

The Schirmer test is the only “objective” measure among the validated scales, and this test has been recommended to measure response in patients with ocular GVHD by the NIH consensus group. Our results, however, suggest that changes in the Schirmer score do not reflect patient-reported changes in ocular GVHD symptoms. The better correlation with clinician-reported changes might occur if clinicians consider results of Schirmer tests when rating clinician-reported symptom changes. Several studies have questioned the intertest reproducibility of the Schirmer test, particularly in patients with mild ocular symptoms.^{34,35} One study concluded that assessment of patient-reported symptoms had greater reliability than the Schirmer test.³⁵ Schirmer test results were not available for 37% of visits, particularly in patients with higher NIH eye scores, attesting to the reluctance of patients or clinicians to conduct this test and reflecting the limited resources to obtain this test in routine clinical care. In 1 study, Schirmer testing in patients with chronic GVHD took 9 minutes,³⁶ required special paper strips, and in some practices, would require an ophthalmology referral. Taken together, Schirmer testing is not recommended to measure the change in activity of ocular GVHD in general chronic GVHD studies.

Our recommendations apply primarily to studies focused on broad populations of patients with chronic GVHD manifestations. For clinical trials specifically targeting ocular symptoms, other detailed and objective ophthalmology criteria might be necessary to document ocular chronic GVHD activity. Candidate measures for validation in future studies include tear breakup time, tear evaporation rate, slit-lamp examination, and corneal staining.³⁷ Because we were specifically evaluating the NIH Consensus Conference recommendations, these measures were not collected in our study.

Our results indicated that agreement in ocular involvement and reported symptom changes between the clinician and patient was only fair. Many explanations are possible. Clinicians and patients may use different criteria to judge ocular GVHD severity, one rater may be more sensitive to symptoms than the other, or 1 or both raters may be poorly calibrated. Because it is not known whose judgment should guide treatment, the ideal scales would correlate with both clinician- and patient-reported symptom changes.

Agreement between clinician and patient ratings of eye involvement and change in eye symptoms was fair. Nevertheless, our results support the use of the NIH eye score to assess response in clinical trials for chronic GVHD for 3 reasons. First, the NIH eye score correlates well with both clinician- and patient-reported symptom changes. Second, the change in this scale shows the greatest sensitivity to reported symptom change. Finally, the NIH eye score takes <15 seconds to complete during an outpatient clinic visit, without requiring a formal patient survey, special equipment, or trained personnel. The use of the NIH eye score alone would dramatically decrease the time required to assess ocular chronic GVHD and would eliminate the need for the Schirmer test.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported in part by grants CA118953 and CA18029 from the National Cancer Institute. Y.I. is a recipient of the Banyu Fellowship Program from Banyu Life Science Foundation International and the JSPS Postdoctoral Fellowships for Research Abroad. The sponsor or funding organization had no role in the design or conduct of this research.

Financial Disclosure(s): The authors have no proprietary or commercial interest in any materials discussed in this article.

References

1. Appelbaum, FR.; Forman, SJ.; Negrin, RS.; Blume, KG. Thomas' hematopoietic cell transplantation. 4th ed.. Wiley-Blackwell; Oxford, England: 2009. p. 705-1199.
2. Socie G, Stone JV, Wingard JR, et al. Late Effects Working Committee of the International Bone Marrow Transplant Registry. Long-term survival and late deaths after allogeneic bone marrow transplantation. *N Engl J Med*. 1999; 341:14–21. [PubMed: 10387937]
3. Lee SJ, Fairclough D, Parsons SK, et al. Recovery after stem-cell transplantation for hematologic diseases. *J Clin Oncol*. 2001; 19:242–52. [PubMed: 11134219]
4. Kiss TL, Abdoell M, Jamal N, et al. Long-term medical outcomes and quality-of-life assessment of patients with chronic myeloid leukemia followed at least 10 years after allogeneic bone marrow transplantation. *J Clin Oncol*. 2002; 20:2334–43. [PubMed: 11981005]
5. Fraser CJ, Bhatia S, Ness K, et al. Impact of chronic graft-versus-host disease on the health status of hematopoietic cell transplantation survivors: a report from the Bone Marrow Transplant Survivor Study. *Blood*. 2006; 108:2867–73. [PubMed: 16788100]
6. Pidala J, Anasetti C, Jim H. Quality of life after allogeneic hematopoietic cell transplantation. *Blood*. 2009; 114:7–19. [PubMed: 19336756]
7. Pidala J, Kurland B, Chai X, et al. Patient reported quality of life is associated with severity of chronic graft-versus-host disease as measured by NIH criteria: report on baseline data from the Chronic GVHD Consortium. *Blood*. 2011; 117:4651–7. [PubMed: 21355084]
8. Lee SJ, Vogelsang G, Flowers ME. Chronic graft-versus-host disease. *Biol Blood Marrow Transplant*. 2003; 9:215–33. [PubMed: 12720215]
9. Calissendorff B, el Azazi M, Lonnqvist B. Dry eye syndrome in long-term follow-up of bone marrow transplanted patients. *Bone Marrow Transplant*. 1989; 4:675–8. [PubMed: 2819286]
10. Tichelli A, Duell T, Weiss M, et al. European Group on Blood and Marrow Transplantation (EBMT) Working Party on Late Effects. Late-onset keratoconjunctivitis sicca syndrome after bone marrow transplantation: incidence and risk factors. *Bone Marrow Transplant*. 1996; 17:1105–11. [PubMed: 8807122]
11. Flowers ME, Parker PM, Johnston LJ, et al. Comparison of chronic graft-versus-host disease after transplantation of peripheral blood stem cells versus bone marrow in allogeneic recipients: long-term follow-up of a randomized trial. *Blood*. 2002; 100:415–9. [PubMed: 12091330]

12. Couriel D, Carpenter PA, Cutler C, et al. Ancillary Therapy and Supportive Care Working Group report. Ancillary therapy and supportive care of chronic graft-versus-host disease: National Institutes of Health consensus development project on criteria for clinical trials in chronic graft-versus-host disease: V. *Biol Blood Marrow Transplant*. 2006; 12:375–96. [PubMed: 16545722]
13. Wolff D, Gerbitz A, Ayuk F, et al. Consensus conference on clinical practice in chronic graft-versus-host disease (GVHD): first-line and topical treatment of chronic GVHD. *Biol Blood Marrow Transplant*. 2010; 16:1611–28. [PubMed: 20601036]
14. Filipovich AH, Weisdorf D, Pavletic S, et al. Diagnosis and Staging Working Group report. National Institutes of Health consensus development project on criteria for clinical trials in chronic graft-versus-host disease: I. *Biol Blood Marrow Transplant*. 2005; 11:945–56. [PubMed: 16338616]
15. Pavletic SZ, Martin P, Lee SJ, et al. Response Criteria Working Group report. 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. *Biol Blood Marrow Transplant*. 2006; 12:252–66. [PubMed: 16503494]
16. Revicki DA, Osoba D, Fairclough D, et al. Recommendations on health-related quality of life research to support labeling and promotional claims in the United States. *Qual Life Res*. 2000; 9:887–900. [PubMed: 11284208]
17. Acquadro, C.; Berzon, R.; Dubois, D.; et al. PRO Harmonization Group. Value Health; Incorporating the patient's perspective into drug development and communication: an ad hoc task force report of the Patient-Reported Outcomes (PRO) Harmonization Group meeting at the Food and Drug Administration; February 16, 2001; 2003. p. 522-31.
18. Cleeland CS, Mendoza TR, Wang XS, et al. Assessing symptom distress in cancer patients: the M.D. Anderson Symptom Inventory. *Cancer*. 2000; 89:1634–46. [PubMed: 11013380]
19. 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; 8:444–52. [PubMed: 12234170]
20. Vitali C, Bombardieri S, Jonsson R, et al. Classification criteria for Sjogren's syndrome: a revised version of the European criteria proposed by the American-European Consensus Group. European Study Group on Classification Criteria for Sjögren's Syndrome. *Ann Rheum Dis*. 2002; 61:554–8. [PubMed: 12006334]
21. Guyatt GH, Osoba D, Wu AW, et al. Methods to explain the clinical significance of health status measures. *Mayo Clin Proc*. 2002; 77:371–83. [PubMed: 11936935]
22. Chronic GVHD Consortium. Rationale and design of the Chronic GVHD Cohort Study: improving outcomes assessment in chronic GVHD. *Biol Blood Marrow Transplant*. 2011; 17:1114–20. [PubMed: 21664473]
23. Schiffman RM, Christianson MD, Jacobsen G, et al. Reliability and validity of the Ocular Surface Disease Index. *Arch Ophthalmol*. 2000; 118:615–21. [PubMed: 10815152]
24. Osoba D, Rodrigues G, Myles J, et al. Interpreting the significance of changes in health-related quality-of-life scores. *J Clin Oncol*. 1998; 16:139–44. [PubMed: 9440735]
25. ASBMT Chronic GvHD Consensus Project. [Accessed August 1, 2011] Measurement of therapeutic response in chronic graft-versus-host-disease. Response Criteria APPENDICES C & D. Available at: http://asbmt.affiniscape.com/associations/11741/files/ResponseCriteriaAPPENDIXC_DCalculations.pdf
26. Norman GR, Sloan JA, Wyrwich KW. Interpretation of changes in health-related quality of life: the remarkable universality of half a standard deviation. *Med Care*. 2003; 41:582–92. [PubMed: 12719681]
27. Norman GR, Sloan JA, Wyrwich KW. Is it simple or simplistic? *Med Care*. 2003; 41:599–600. [PubMed: 12719684]
28. Fitzmaurice, GM.; Laird, NM.; Ware, JH. Applied Longitudinal Analysis. Wiley; Hoboken, NJ: 2004. p. 325-58.
29. Gardiner JC, Luo Z, Roman LA. Fixed effects, random effects and GEE: what are the differences? *Stat Med*. 2009; 28:221–39. [PubMed: 19012297]

30. Davitt WF, Bloomenstein M, Christensen M, Martin AE. Efficacy in patients with dry eye after treatment with a new lubricant eye drop formulation. *J Ocul Pharmacol Ther.* 2010; 26:347–53. [PubMed: 20653478]
31. Friedman NJ. Impact of dry eye disease and treatment on quality of life. *Curr Opin Ophthalmol.* 2010; 21:310–6. [PubMed: 20467319]
32. Koffler BH, McDonald M, Nelinson DS, LAC-07-01 Study Group. Improved signs, symptoms, and quality of life associated with dry eye syndrome: hydroxypropyl cellulose ophthalmic insert patient registry. *Eye Contact Lens.* 2010; 36:170–6. [PubMed: 20351555]
33. Takahide K, Parker PM, Wu M, et al. Use of fluid-ventilated, gas-permeable scleral lens for management of severe keratoconjunctivitis sicca secondary to chronic graft-versus-host disease. *Biol Blood Marrow Transplant.* 2007; 13:1016–21. [PubMed: 17697963]
34. Lee JH, Hyun PM. The reproducibility of the Schirmer test. *Korean J Ophthalmol.* 1988; 2:5–8. [PubMed: 3079546]
35. Nichols KK, Mitchell GL, Zadnik K. The repeatability of clinical measurements of dry eye. *Cornea.* 2004; 23:272–85. [PubMed: 15084861]
36. Mitchell SA, Jacobsohn D, Thormann Powers KE, et al. A multicenter pilot evaluation of the National Institutes of Health chronic graft-versus-host disease (cGVHD) therapeutic response measures: feasibility, interrater reliability, and minimum detectable change. *Biol Blood Marrow Transplant.* 2011; 17:1619–29. [PubMed: 21536143]
37. Wang Y, Ogawa Y, Dogru M, et al. Baseline profiles of ocular surface and tear dynamics after allogeneic hematopoietic stem cell transplantation in patients with or without chronic GVHD-related dry eye. *Bone Marrow Transplant.* 2010; 45:1077–83. [PubMed: 19898506]

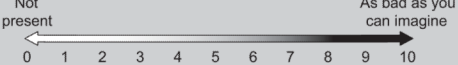
Measurement scale	No. of items	Score
National Institutes of Health (NIH) eye score	1	(0) No symptoms (1) Mild dry eye symptoms not affecting ADL (requiring eye drops $\leq 3 \times$ per day) OR asymptomatic signs of keratoconjunctivitis sicca (2) Moderate dry eye symptoms partially affecting ADL (requiring eye drops $> 3 \times$ per day or punctal plugs) WITHOUT vision impairment (3) Severe dry eye symptoms significantly affecting ADL (special eyewear to relieve pain) OR unable to work because of ocular symptoms OR loss of vision caused by keratoconjunctivitis sicca
Patient-reported global rating of eye symptoms	1	Please circle the number that shows how severe your symptoms have been in the last week : Your eye problem at its WORST? Not present As bad as you can imagine 
Eye subscale of Lee symptom scale	3	Summary of the following 3 items (0 – 100) -Dry eyes -Need to use eye drops frequently -Difficulty seeing clearly Not at all Slightly Moderately Quite a bit Extremely 0 1 2 3 4
Ocular surface disease index (OSDI)	12	Summary of 12 items (0 – 100)
Schirmer test	1	mm in 5 minutes
Clinician or patient-reported symptom change ("Anchor")		(1) Completely gone (2) Very much better (3) Moderately better (4) A little better (5) About the same (6) A little worse (7) Moderately worse (8) Very much worse

Figure 1. Measurement scales validated and an 8-point scale of reported symptom change. ADL = activity of daily living.

A

		Patient-reported involvement	
		No	Yes
Clinician-reported involvement	No	83 (25%)	77 (23%)
	Yes	21 (6%)	154 (46%)

Kappa=0.40 (fair agreement)

B

		Patient-reported symptom change		
		Improvement	Stable	Worsening
Clinician-reported symptom change	Improvement	50 (18%)	27 (10%)	5 (2%)
	Stable	49 (17%)	112 (40%)	24 (9%)
	Worsening	0 (0%)	6 (2%)	4 (1%)

Fleiss-Cohen weighed kappa=0.34 (fair agreement)

Figure 2.

Agreement between the clinician and the patient in **(A)** eye involvement at enrollment, and **(B)** reported symptom change at the follow-up visit. **A**, n = 335. Fifty-two of 387 patients (13%) missed eye measurement scales at enrollment. **B**, n = 277. Fifty-six of 333 follow-up visits (17%) with both clinician and patient reporting eye involvement at the previous visit were missing clinician or patient-reported symptom change.

Agreement in response between the measurement scales (kappa statistic)

	National Institutes of Health (NIH) eye score	Global rating of eye symptoms	Lee eye subscale	Ocular surface disease index
Global rating of eye symptoms	0.18 (slight)			
Lee eye subscale	0.26 (fair)	0.42 (moderate)		
Ocular surface disease index	0.14 (slight)	0.37 (fair)	0.47 (moderate)	
Schirmer test	0.18 (slight)	0.10 (slight)	0.03 (slight)	-0.02 (no)

Figure 4.

Agreement in response between the measurement scales (kappa statistic). The kappa value and its interpretation are shown for each comparison. Response (improvement, stable, or worsening) was based on clinically meaningful change for each scale.

Table 2
Correlation of Changes in Measurement Scales with Symptom Changes Reported by Clinicians or Patients

Measurement Scale	Clinically Meaningful Change in Scale*	Visits with Clinically Meaningful Change (%)	Clinician-Reported Symptom Change		Patient-Reported Symptom Change		P	
			n	Estimate [†] (95% CI)	P	n		Estimate [‡] (95% CI)
NIH eye score	1 point	23% Improve 15% Worsen	300	1.26 (1.01–1.51)	<.001	268	0.43 (0.13–0.73)	0.005
Patient-reported global rating of eye symptoms	2 points	23% Improve 25% Worsen	256	0.17 (0.01–0.33)	.03	259	0.35 (0.21–0.49)	<0.001
Eye subscale of Lee symptom scale	15 points	29% Improve 26% Worsen	263	0.19 (0.06–0.33)	.004	265	0.29 (0.17–0.41)	<.001
Ocular surface disease index	10 points	24% Improve 34% Worsen	250	0.11 (0.01–0.21)	.04	253	0.19 (0.10–0.28)	<0.001
Schirmer test [‡]	–5 mm	22% Improve 22% Worsen	128	0.23 (0.03–0.42)	.02	118	0.06 (–0.10–0.22)	0.47

CI = confidence interval; NIH = National Institutes of Health.

* See Methods for definition.

[†]The estimates of reported symptom change on an 8-point scale (Fig 1) were modeled according to units originally defined as clinically meaningful for each scale. For example, a 15-point increase in the Lee eye subscale correlated with a 0.19-point worsening on the 8-point clinician-reported scale, and a 5-mm decrease in the Schirmer test correlated with a 0.06 worsening on the 8-point patient-reported scale. All models controlled for type of enrollment (incident versus prevalent).

[‡]The value of the worst eye.